ANALYSIS OF CENTRAL VENOUS LINE RELATED THROMBOEMBOLISM AND INFECTION IN CHILDREN WITH ACUTE LYMPHOBLASTIC LEUKEMIA

EXPLORATORY ANALYSIS TO DETERMINE PREVALENCES AND PREDICTORS OF CENTRAL VENOUS LINE RELATED THROMBOEMBOLISM AND INFECTION IN CHILDREN WITH ACUTE LYMPHOBLASTIC LEUKEMIA

By

WEIWEI XIONG, B.SC, MBA

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2008) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Exploratory Analysis to Determine Prevalences and Predictors of Central Venous Line Related Thrmoboembolism and Infection in Chidren with Acute Lymphoblastic Leukemia

AUTHOR: Weiwei Xiong

B.Sc. (Wuhan University, China)

MBA (Fudan University, China)

SUPERVISOR: Dr. Lehana Thabane

NUMBER OF PAGES: x, 96

ABSTRACT

Children with acute lymphoblastic leukemia (ALL) are at high risk for getting thromboembolism (TE), which is a serious complication leading to morbidity and mortality. As treatment protocols have been developed achieving the cure rates as high as 80% [39], study efforts need to turning to evaluating the risk and management of associated TE. Published studies in this field have been mostly exploratory and have had different results in terms of screening TE risk factors predisposing to TE.

Based on the records of 150 ALL children treated with central venous line (CVL) from 1995 to 2005 at McMaster Children's Hospital, this study was conducted to evaluate the prevalence of TE, to explore the association between TE and infection, and to screen TE and Infection risk factors disposing children with ALL for TE and for Infection. The prevalence of TE was estimated as 15.07% (9.27%, 20.87%). Logistic regressions, Bayesian approaches, in combination with multiple imputation techniques, were employed to estimate predictors' odds ratios and their 95% confidence (credibility) intervals. The study suggested two significant factors, CVL functionality and ANC category for infection, and no significant factors for TE.

As a comparative and supplementary tool to the traditional parametric analyses, we conducted Classification and Regression Trees (CART) modeling, by using three software packages, with intention to visualize predictors of TE and Infection by level of importance. SAS EM 5.0, SPSS 14.0 and S-Plus 6.1 were compared in their tree misclassifications based on our data and their features of tree growth algorithms, validation techniques, missing data handling, model pruning / recovering, output setting, tool tabs transparency, and advantages. SPSS 14.0 and SAS EM 5.0 are recommended based on our experience, though the strengths and weaknesses of each package should be weighted according to the users and the problem natures.

The limitations of this exploratory study such as small sample size, missing values, imbalance between data categories, the lack of information about the timing of treatment and the lack of cross-validation techniques in some CART modeling packages

led biases to our results. Large prospective cohort studies with few missing values are critical to achieve more accurate results.

ACKNOWLEDGEMENT

I would like to express my gratitude to my project supervisor, Dr. Lehana Thabane, for his instruction, guidance, encouragement and patience throughout this study. I appreciate the time and opportunities he arranged for me to participate in health methodological rounds and to improve my presentation skills, which turned out to be very valuable.

I am thankful to Dr. Uma Athale for providing the data, reading my report and providing constructive comments. Thanks to Anita Lathia for collecting these data.

I am grateful to Dr. Roman Viveros, Dr. N. Balakrishnan, Dr. Peter Macdonald, and Dr. Angelo Canty for their instructions and advices. Their encouragement and understanding is vital for me to complete my M.Sc study.

This thesis is in memory of my beloved mom, and to my dad, to whom I am deeply indebted for their love and for giving me life. I am also grateful to my sister and brother-in-law for taking care of dad on my behalf.

To my friends Yan Chen, her husband Ming Wei and Jinhui Ma, I appreciate their moral encouragement and consistent caring.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODCTION

1.1 Overview of the Problem

Acute lymphoblastic leukemia (ALL) is a common childhood disease. Being associated with ALL, thromboembolism (TE) is recognized as a serious complication leading to significant morbidity. The reported prevalence of TE in children with ALL varies from the lowest 1.1% to the highest 36.7% [1, 2, 3], compared to the much lower estimates of the prevalence of deep venous thrombosis (DVT) and pulmonary embolism in the general pediatric population and hospital admissions [4, 5, 6].

The occurrence of TE seems to be emerging from the interaction of the disease with the therapy and possible genetic predisposition for hypercoagulability [1]. Compared to adult malignancy-related TE, on which many studies have been conducted and many evidence-based guidelines have been established, quantitative studies on ALL-related TE in children are limited, and have inconsistent results.

Central Venous Line (CVL) or Central Venous Catheter is a commonly used catheter inserted into a large vein in the neck (jugular vein), chest (subclavian vein) or groin (femoral vein) for chemotherapy. It is a recognized risk factor for infection. However, little is known about the association between infection and TE, or impact of any optimal techniques and locations for CVL insertion, if they exist.

1.2 Objectives of the Thesis

Based on a non-interventional retrospective study of 150 treatment cases at McMaster Children's Hospital, this thesis aims to tackle the following clinical and statistical objectives:

1.2.1 Clinical Objectives

The first part of the thesis addresses the original clinical objectives of the study, which were as follows:

a. To determine the prevalence of TE and the prevalence of infection in children with ALL over the period from 1995 to 2005;
b. To detect association between TE and infection in children with ALL; and
c. To generate hypotheses about risk factors for TE and for infection in children with ALL.

1.2.2 Statistical Objectives

The second part of the thesis objectives involved some statistical issues. In particular, the statistical objectives were:

a. To generate hypotheses about important risk factors of CVL-related TE and infection;
b. To compare various analyses targeting on clinical objectives b and c. The approaches are
    a) Simple logistic regression without multiple imputation (MI) [7]
    b) Fisher's exact test without multiple imputation
    c) Logistic regression with rounding incorporated MI [7, 8, 9,10,11]
    d) Logistic regression with non-rounding incorporated MI [12, 13]
    e) Bayesian analysis [17, 18, 19]
    f) Classification and Regression Trees (CART) modeling [23, 24] without MI
    g) CART modeling with MI
b. To construct by clinical importance ordered trees for TE and Infection for therapeutic and preventive purpose;
c. To compare SAS, SPSS and S-Plus in CART modeling.

1.3 Ethical Considerations

This study was conducted in accordance with the Trial Council Policy Statement Guidelines [25], including the Good Clinical Practice (GCP) Guidelines [25].

The study was approved by the McMaster University Research Ethics Board. Since

it is a retrospective chart review, it does not impose any additional risk to the patients.

1.4 Significance of the Study

As more aggressive treatment protocols are being developed to achieve high cure rates, studies evaluating the risk and management of associated TE have become increasingly important. Being a widely used instrument in ALL therapy, CVL is naturally suspected in association with increased risk of TE. To prevent TE-related morbidity, we need to identify patients at risk of TE first and then move on to develop therapeutic protocols and guidelines in CVL-involved disease treatment.

Although the conclusions drawn from this study are based on data from patients with ALL, it is rational to believe that similar mechanisms exist with other CVL-involving treatments. Greater knowledge of how the site and technique of CVL insertion affect TE is essential in setting up future management guidelines for CVL.

The results of this study will be useful in designing further studies to definitively explore the risk factors of CVL-related TE and infection, not only in patients with ALL, but also in patients with medical conditions that require CVL treatment. Using different statistical methods will provide useful information about the robustness of the findings.

1.5 Scope of the Report

Limited by a small sample size of 150, a small event (TE) percentage (22 out of 150), and a noticeable missing data rate, the presented results of the study are exploratory. By conducting this study, we intended to generate hypotheses for future research.

Chapter 2 covers analysis methods used in this study. We first estimate the prevalence of TE and the prevalence of infection and their 95% confidence intervals in children with ALL, from 1995 to 2005. Then, based on patient demographics and pre-screened factors for TE and infection, we discuss three logistic regression analyses with

and without MI, in combination with rounding and non-rounding considerations. We also describe Bayesian analysis approach.

Next, we describe CART modeling as a different approach from traditional parametric analysis, with and without MI integrated. The aim was to investigate how the results of CART modeling were different from those of logistic and Bayesian analyses. SAS EM 5.0, SPSS 14.0 and S-Plus 6.1 Tree Functions are discussed with respective built-in node splitting criterion. The aim is to compare the three types of software for implementing CART modeling and to construct clinical importance ordered trees for TE and Infection for therapeutic and preventive strategies.

Chapter 3 covers all results created by approaches discussed in Chapter 2.

Chapter 4 provides some discussions of the results, including hypotheses generated by results in Chapter 3. We discuss limitations of this study and give suggestions for future studies.

Chapter 5 provides some concluding remarks, focusing on the key messages of the thesis.

2 STUDY METHODS

2.1 Description of the Data Set

This study is a retrospective chart review. We retrieved 150 records of children (≤ 18 years of age) who were treated with CVL at McMaster Children's Hospital from 1995 to 2005. Each record consists of 25 variables, of which five are basic patient information, nine about CVL, six about infection, two about TE and the other three about asparaginase, linogram and outcome of death.

For those cases with one or more factor values changing in the course of treatments, such as Type of CVL, CVL Insertion Technique and Site of CVL Insertion, we retrieved their first-time values and ignored those at later times so as to avoid possible within-patient correlation issues.

2.2 Statistical Analyses

2.2.1 Description of Outcomes and Predictors

We chose 10 out of 25 variables for analysis in our study.

TE and Infection were taken as outcome or response variables. The predictors considered for TE were Infection, Age, Gender, ALL-Risk category, Phase of Chemotherapy at CVL Insertion, CVL Insertion Technique, CVL Functionality, Type of Asparaginase, Type of CVL and Site of CVL Insertion. The predictors considered for Infection were TE, Age, Gender, ALL-Risk category, Phase of Chemotherapy at CVL Insertion, CVL Insertion Technique, CVL Functionality, Type of CVL, Site of CVL Insertion and Absolute Neutrophil Count (ANC).

2.2.2 Description of Outcomes Including Coding

Both deep venous thrombosis (DVT) and pulmonary embolism (PE) cases were counted as TE.

Infection cases were screened with positive CVL blood culture readings.

2.2.3 Description of Predictors Including Coding

Tables 2.5 and 2.6 break down values of all the outcome and predictor variables in the form of value, count, percentage and coding scheme. Variables renamed in short forms are listed in the first column, with their full names included in brackets.

Age and ANC data were originally ordinal or continuous. They were redefined as binary variables depending on the coverage in which their continuous counterparts fall. Both the analyses with continuous variables, and the analyses with categorical variables were conducted.

Category	→	**Continuous Value Coverage**
Age category 1	→	≥ 10 years of age
Age category 2	→	< 10 years of age
ANC category 1	→	$< 0.5\ 10^9/L$
ANC category 2	→	$\geq 0.5\ 10^9/L$

ALL-Risk has two categorical values: Standard Risk (SR) and High Risk (HR). By definition, SR includes cases satisfying: (1) age of 1 to 10 years; (2) precursor B Cell type of ALL; (3) absent mediastinal mass; (4) white blood cell (WBC) reading less than 50K; (5) absent central nervous system (CNS) ALL. Cases not satisfying any of the above conditions belong to HR.

2.2.4 Imbalance of the Data

Tables 2.3 and 2.4 provide patient demographics and prognostics by TE and Infection.

Four records have TE status missing and five records have Infection status missing. We observed that apart from Age, ANC and binary variables Gender, ALL-Risk, CVL Insertion Technique, CVL Functionality, Site of CVL Insertion, three multi-category variables presented data imbalance. For Chemotherapy Phase at CVL Insertion, seven cases dispersed into 5 categories. We grouped them into the "Post Chemotherapy" category, in parallel to other two dominant groups "Prior to Chemotherapy" and "Induction". For Type of Asparaginase, out of 4 groups, only

Escherichiae Coli (E.Coli) have both present and absent TE cases. Similarly, Type of CVL consists of 140 cases in one single group, but only six cases in the other two groups. After removing these two variables, we ended up with a dataset of 2 continuous, 1 ternary and 7 binary variables for logistic and Bayesian analyses. We included Type of CVL and Type of Asparaginase in Fisher's exact tests.

2.3 Missing Data and Imputation

For TE, fifteen records (10%) have at least one variable with missing values. For Infection, twenty four records (16%) have at least one variable with missing values. We need to take into account these considerable missing rates in order to avoid further power loss of our study.

Many methods are available for handling missing data. The simplest approach is to ignore missing data records and to do complete-case multivariate analysis or available-case univariate analysis. Such kinds of analysis, based on completely observed values, result in decreased power due to sample size reduction. This makes difference especially for frequently missing or small sample size data. In our study, we conducted available-case analysis by logistic Bayesian approaches.

Single imputation procedures such as mean imputation, hot-deck / cold-deck imputation, regression imputation, stochastic regression imputation and last observation carried forward, composite method etc., replace each missing value with "an" estimate by some certain rules [27, 28]. They do not take into account the sampling variability produced by the imputed values, hence they generally result in underestimation of the variance [27, 28].

Multiple imputation (MI) procedure [8-11], in contrast, is a simulation-based approach which replaces each missing value with a set of m plausible values by assuming that all involved variables in the imputation are multivariate normally distributed. Each of the m created complete versions of the data set are analyzed by a pre-specified method. Results of point estimates and estimated standard error are generated as usual. Then by Rubin's rule [9, 11], they are pooled to arrive at a final single result of the point

estimate and associated confidence interval. The MI approach is advantageous over other imputation approaches in terms of minimizing inference bias, maintaining power, and reflecting sampling variability [7-11]. It serves as the optimal approach to missing values in most scenarios.

SAS Proc MI and SAS MIANALYZE [14, 15, 16] were employed to conduct MI analysis. In particular, the Markov Chain Monte Carlo simulation (MCMC) technique was used instead of the other two built-in techniques (parametric regression method and non-parametric propensity scores method) based on the arbitrary missing-at-random (MAR) pattern of the data being studied. However, the multivariate normality assumption on which MCMC is based is not exactly satisfied because most of our variables with missing values are not continuous but categorical. Should we round the imputed values to their closest categorical values so that all created m datasets preserve the same categorical frame as the original but with multivariate normality being compromised somehow, or keep the imputed values as they are with the m datasets not strictly preserving categorical data pattern? Simulation studies [12] show that rounding in multiple imputation leads to biasness. They suggest retaining imputed values non-rounded for analysis [13].

To "fill in" the missing values, we employed a multiple imputation approach to create five imputed datasets and took rounding and non-rounding strategies separately. These two strategies incorporated to the logistic model led to our second and third logistic approaches. Five non-rounded imputed data sets and five corresponding imputed datasets with values rounded to their closest categorical values were saved for separate logistic analyses. Results drawn from original data, multiple imputed rounded data and multiple imputed non-rounded data are to be compared.

2.4 Brief Description of the Different Statistical Methods

2.4.1 Outcome Prevalence Estimates

Prevalence of TE (or infection) was estimated based on that the probability that

for each ALL child treated with CVL to have TE complication (or infection) is identically independently Bernoulli distributed. In other words, the number of CVL-related TE (or infection) in ALL children follows Binomial distribution. By computing the proportion of TE (or infection) over the number of available cases, which is 146 (TE) or 145 (infection), we estimated 1995-2005 TE and infection prevalence *p's*. The standard deviation *sd's* of the estimate p's were approximated by

$$sd = \sqrt{p(1-p)/n}$$

The 95% confidence intervals for the estimates of TE and infection prevalence were approximated as

$$(p-1.96sd,\ p+1.96sd)$$

2.4.2 Logistic Regressions

To generate hypotheses about risk factors of TE and Infection, Simple logistic regressions with available cases were conducted first, then logistic regressions with rounded / non-rounded multiple imputation approaches were conducted. The model is shown below.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i \qquad i = 1, 2, \ldots, n$$

p_i = Probability ($y_i = 1$)

where

x_i being the ith sampled value of the predictor variable;

y_i being the ith sampled value of the outcome variable; and

α, β being the parameter estimates.

The numbers of TE and Infection are 22 (15.07%) and 83 (57.24%). This limits to some extent the number of covariates being analyzed simultaneously in logistic regression especially for outcome TE. Starting from univariate regression, and by using forward stepwise selection, we aimed to find as many significant factors as appropriate

(limited to 2 for TE and 8 for Infection). In a study like ours with small sample size and low event incurrence (for TE), one can only conduct exploratory analysis and generate hypotheses for further verification by future researches.

2.4.3 Bayesian Analysis

Next, we employed Winbugs 14 [22] to conduct the Bayesian analysis on the original dataset based on the following model.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i \qquad i = 1, 2, \ldots, n$$

p_i = Probability ($y_i = 1$)

where

x_i being the ith sampled value of the predictor variable;

y_i being the ith sampled value of the outcome variable; and

α, β being the parameter which follow certain distributions.

We conducted the Bayesian analysis by introducing a non-informative normal prior distribution $N(0, 0.0001)$ for α and β, by setting 3000 burn-in's and 10000 iterations, aiming to achieve converged estimates of β's (95% credibility intervals). Note that for WinBugs one needs to state the precision for the Normal prior instead of the variance. Bayesian analysis results are then to be compared with those resulted from classical logistic regression approaches.

The reporting of the Bayesian results is done in accordance with the ROBUST (Reporting Of Bayes Used in clinical STudies) guidelines for Bayesian Analysis of Clinical Studies. [21]

2.4.4 CART Modeling

While parametric approaches such as logistic and Bayesian analyses are widely used in estimating regression parameters and thus screening significant factors for the outcomes, CART modeling uses efficient non-parametric data mining algorithms to

generate binary or k-nary trees, along which predictors are top-down ordered by their importance in predicting the outcome.

In our study, CART modeling was conducted with the same data set by using three software packages SAS EM 5.0, SPSS 14.0 and S-Plus 6.1. We intended to compare difference of traditional regression approaches and CART modeling in predicting outcomes. As various tree growth and node splitting rules are integrated with each software package, we also intended to compare the above three packages in their predicting misclassification rates and their CART modeling functions.

Table 2.2 provides a summary of three main tree growth and node splitting rules. They are Automated Interaction Detection (CHAID) [29], Classification and Regression Tree (CART / CRT) [24] and Quick Unbiased Efficient Statistical Tree (QUEST) [30].

For CART / CRT, the split is selected which divides the observations at a node into subgroups in which a single class predominates. The tree reaches a leaf node until no split can be found that increases the class specificity at a node. When all observations are in leaf nodes the tree has stopped growing. Each leaf can then be assigned a class and an error rate. The tree may be cut back to a size which allows effective generalization to new data. Branches of the tree that do not enhance predictive classification accuracy are eliminated in a process known as "pruning". Three main impurity reduction criteria employed by CART / CRT are Entropy reduction, Gini-index and Twoing [24]. Table 2.2 provides their related purity functions.

CHAID differs from CART in that it stops growing a tree before over-fitting occurs. When there are no more splits available that lead to a statistically significant improvement in classification, the tree stops growing. By using CHAID, any continuously valued attributes must be redone as categorical variables. Chi-Square-tests (Pearson / likelihood ratio), with / without Bonferroni adjustments are the common criteria employed by CHAID. [29]

QUEST is another type of decision tree which performs approximately unbiasedness as to class membership variable selection to split nodes. It separates splitting predicator selection into variable selection by F-test and Chi-Square-test (with or

without Bonferroni adjustment) and split point selection by quadratic discriminate analysis (QDA). [30]

First, In order for the results of CART modeling to be comparable to the results of the traditional approaches, we include the same predictors as before. For age and ANC, we take their categorical variables Agee_cate and ANC_cate, rather than their continuous counterparts Age_cont and ANC_cont, in the CART modeling.

Considering the difference between TE and Infection in their event rates and numbers of significant predictors by traditional ways, we set the maximum depth of tree for TE and for Infection as 3 and 5 respectively. The other basic options for both were set as: (1) maximum number of branches from a node: 2; (2) minimum number of observations in a leaf: 5; (3) observations required for a split search: 10; (4) number of candidate rules saved in a node: 5; (5) surrogate rules saved in each node: 5; (6) significance level (CHAID and QUEST): 20%; (7) treat missing as an acceptable value.

As the three software packages differ in validation [32, 33, 34], and the study has a small sample size, we chose not to separate the data into groups of training, validation and testing whenever possible (SPSS 14.0, S-Plus 6.1) [32, 33, 34], to include as many records as possible (SAS 5.0) [32, 33, 34] and to employ 10-fold cross-validation [31] whenever possible (SPSS 14.0) [32, 33, 34] in estimating the misclassification. Misclassification rates resulting from the original data are to be compared to the averaged misclassification rates of the five "filled-in" data sets by multiple imputations.

Secondly, to make the generated trees for TE and Infection more helpful for therapeutic and preventive strategies, we conducted CART modeling differently by selecting six predictors interactively in certain time orders. The six predictors time orderly for TE are CVL Insertion Technique, Site of CVL Insertion, Age categorical or Gender, CVL functionality and Infection. The 6 predictors time orderly for Infection are CVL Insertion Technique, Site of CVL Insertion, Age categorical or Gender, CVL functionality and TE.

Finally, based on our experience using the three software packages in conducting CART modeling with the project data, we compare their overall functionality and

feasibility and give our choice recommendation accordingly.

3 RESULTS

3.1 Descriptive Statistics for the Sample

Patient demographics and prognostics by TE and Infection are summarized in Tables 2.3 and 2.4.

3.1.1 Key Demographics

The average age of the patients having TE was 7.27 years with standard deviation 5.22 years, whereas the age averaged on those having no TE was 6.27 years with standard deviation 4.20 years. The average age of the patients being infected was 6.97 years with standard deviation 4.79 years, whereas the age averaged on those not infected was 5.68 years with standard deviation 3.65 years. About 18% (11) and 58% (36) of female patients had TE and were infected respectively. About 13% (11) and 57% (47) of male patients had TE and were infected respectively.

3.1.2 Key Prognostics

About 19% (14) patients who had CVL inserted before induction of chemotherapy, about 11% (7) of patients who had CVL inserted during chemotherapy induction and about 33% (1) of patients who had CVL inserted after induction of chemotherapy developed TE. Respectively, about 58% (41) of those with CVL inserted before induction of chemotherapy, about 60% (40) of those with CVL inserted during induction of chemotherapy and about 29% (2) of those with CVL inserted after induction of chemotherapy were infected. About 21% (7) of patients with CVL dysfunction and about 13% (14) of those with CVL functioning well had TE. 25% (1) of patients who had their CVL functionality information missing had TE. About 74% (26) and about 51% (54) of patients with CVL dysfunction and CVL functioning well were infected. 60% (3) patients with their CVL functionality information missing were infected. About 8% (3) of cases with CVL inserted at the left side of the body had TE, while about 18% (19) of those with CVL inserted at the right side of the body had TE. 9 patients had their absolute

neutrophil count (ANC) data missing. Average ANC readings of infected and not infected cases were 1.97 10^9/L and 1.79 10^9/L with standard deviation 3.64 10^9/L and 2.60 10^9/L respectively.

About 14% (19) and 55% (77) of patients treated with portacath type of CVL had TE and infection respectively. Of the 5 patients treated with Hickman line type of CVL and 1 treated with a peripherally inserted central catheter (PICC), 50% had TE and all were infected. Portacath cases influentially dominated the type of CVL (96%). The data pattern of this variable was very imbalanced. We excluded it in our study. Type of Asparaginase was also excluded for a similar reason.

3.2 Clinical Results

3.2.1 Prevalence of TE and Prevalence of Infection

The prevalence of TE was estimated as 15.07% (9.27%, 20.87%). The prevalence of Infection was estimated as 57.24% (49.19%, 65.29%).

3.2.2 Key Findings on Those Which Are Statistically Significant

Tables 3.1 and 3.2 summarize odds ratio estimates by TE and by Infection, and their 95% confidence (credibility) intervals, using logistic regressions, Fisher's exact test and Bayesian analysis. Figures 3.1 and 3.2 provide Bayesian convergence plots and results by TE and by Infection. Statistically, the results did not reveal any of the 8 predictors significant to TE, but identified CVL functionality (2.728 (1.167, 6.378)) and ANC category (2.180 (1.063, 4.475)) are significant predictors of Infection. Patients with CVL not functioning properly had over 2.7 times the odds of infection compared to patients with CVL functioning properly. Patients with ANC readings below 0.5 had over 2 times the odds of infection compared to patients with ANC above or equal to 0.5.

3.2.3 Key Findings on Those Which Are Not Statistically Significant, but Are Clinically Important

Four predictors, although not statistically significant, were found to be clinically important to TE. They were Infection (1.727 (0.657, 4.538)), Age category (0.541(0.199, 1.469)), CVL functionality (1.741 (0.638, 4.747)) and Site of CVL insertion (0.404 (0.112, 1.454)). In particular, patients being infected, patients either younger than 1 year of age or older than 10 years of age, patients with CVL not functioning properly, and patients with CVL inserted at the right side of the body all had nearly over 2 times the odds of TE compared to those in the opposite categories. However the 95% confidence (credibility) intervals of these odds ratio estimates ranged from below 1 to over 1. They did not show statistical significance.

Three predictors were found not statistically significant but clinically important to Infection. They were TE (1.727 (0.657, 4.538)), Age category (0.533 (0.232, 1.228)) and Phase of chemotherapy at CVL insertion (3.416 (0.620, 18.810) prior to vs. post, 3.703 (0.669, 20.489) induction vs. post). In particular, patients having TE, patients either younger than 1 year of age or older than 10 years of age, patients having CVL being inserted before or during chemotherapy, and patients with ANC less than 0.5 109/L all had nearly over 2 times the odds of infection compared to those in the opposite categories. However the 95% confidence (credibility) intervals of these odds ratio estimates ranged from below 1 to over 1. They did not show statistical significance.

3.3 Statistical Results

3.3.1 Sensitivity Analysis with Different Methods

The point estimates of odds ratios by TE and by Infection resulted from the simple logistic regression, the logistic regression with rounding incorporated and non-rounding incorporated MI, Fisher's exact test, and Bayesian analysis are close to each other with respective to every predictors. The only observed difference is that the odds ratio confidence (credibility) intervals with respect to ANC category estimated by simple

logistic regression (2.180 (1.063, 4.475)) and Bayesian method (2.202 (1.064, 4.513)) suggest their significance to Infection, whereas the logistic regression with rounding incorporated MI (1.974 (0.943, 4.134)) and the logistic regression with non-rounding incorporated MI (1.974 (0.943, 4.131)) didn't imply the significance of ANC category to Infection outcome. This trivial difference doesn't imply real inconsistency because the lower bounds of 95% confidence intervals by MI incorporated logistic regressions, which are 0.943, are very close to 1.

Based on our original data, TE decision tree examples are provided in Figures 3.5, 3.6 and 3.7. They were generated by SAS EM 5.0, SPSS 14.0 and S-Plus 6.1 respectively. Infection decision tree examples generated by the three software packages are provided in Figures 3.8, 3.9 and 3.10. These decision trees show that the predictors by their importance to TE (high to low) include Site of CVL insertion in the first level, CVL functionality and Phase of Chemotherapy at CVL insertion in the second level, Age category, ALL-risk, Gender and Infection in the third level. These are the variables worthy of attention when predicting TE status. Most of them were also screened by traditional analysis (see above) as clinically important to TE. The predictors by their importance to Infection (high to low) include CVL functionality in the first level, ANC category in the second, TE in the third, Site of CVL insertion and Gender in the fourth, and ALL-risk in the fifth. These are the variables worthy of attention when predicting Infection status. The variables in the first three levels were the same as those identified statistically significant or clinically important to Infection by the traditional analyses. The other three in the lower levels of the tree, however, were different with the results of those traditional approaches.

Figure 3.11 presents clinical importance ordered trees for TE and Infection by using SAS EM 5.0 on the original data set. They were generated by interactively selecting among the given six factors in a certain time orders, which were described in 2.4.4. The average misclassification rates of the clinical importance ordered trees in predicting the outcomes were 15.11% and 39.37%, respectively.

Tables 3.3 and 3.4 summarize CART modeling misclassification rates by SAS EM

5.0, SPSS 14.0 and S-Plus 6.1, each based on the data sets of our study, with and without MI. In predicting TE, the misclassification rates resulted from the three packages with the original data set, on average, are 15.11%, 14.86% and 15.79%, respectively. The misclassification rates resulted from the three packages with the MI data sets, on average, are 15.61%, 15.32% and 15.20%, respectively.

In predicting Infection, the misclassification rates resulting from the packages with the original data set, on average, are 41.96%, 35.76% and 29.37% respectively. The misclassification rates resulting from the three packages with the MI data sets, on average, are 38.51%, 35.53% and 32.40% respectively.

Generally, SPSS trees had lower misclassification rates than SAS trees in both TE and Infection predictions, whereas S-Plus trees had the highest misclassification rates in TE cases and the lowest misclassification rates in Infection cases. The misclassification rates with MI data sets in TE predictions were slightly higher than those with the original data set for whichever tree growth methods used, whereas in Infection predictions, the misclassification rates with MI data sets were generally about 3% and 1% lower than those with the original data set. SPSS 10-fold validation and split sampling techniques resulted in consistent averaged misclassification rates. SPSS CHAID misclassifications were close to SAS Chi-Square misclassifications. Its CRT with Gini or Twoing gave consistently lower misclassifications than all other techniques. Its QUEST algorithm gave the highest misclassifications in almost all cases.

3.3.2 Impact of Missingness

Missingness in our study didn't make much difference in terms of odds ratio estimates. Analyses with and without MI gave consistent odds ratio estimates. Likewise, a minor difference existed when calculating 95% confidence intervals of odds ratios by Infection for CVL functionality and ANC category. Analyses with MI led to a statistical insignificance conclusion, with the lower bounds just passing 1, where as simple logistic regression and Bayesian analysis suggested their significance.

As for rounding versus non-rounding incorporated MI, the results were all

consistent.

Figures 3.3 and 3.4 are forest plot examples which show by TE and by Infection the point estimates and 95% confidence (credibility) intervals of particular odds ratios, by using the three different logistic approaches and Bayesian approach.

3.3.3 Comparison of Cart Modeling Packages

Table 3.5 summarizes CART modeling features built in SAS EM 5.0, SPSS 14.0 and S-Plus 6.1 including tree growth , validation technique, missing data handling, model pruning and recovering, output setting, tool tabs transparency, and advantages.

SAS EM 5.0 provides a built-in comparison between CART modeling and logistic regression. It supports both automatic and interactive training of the model, at each level with each node. Tree pruning, recovering and node selecting are a matter of a click at the users' discretion. Data set, attribute settings, partitions, and tree options are connected as a project, and they can conveniently be adjusted and saved for modeling with a switched data set. However, SAS EM 5.0 requires the sample to be separated into training, validation and test, no matter how small the sample size might already be. Misclassification is assessed based on the training group or combined group of training and validation.

With SPSS 14.0, sample grouping is an option, not a must, for misclassification assessment. Validation can be none, split sampling or V-fold cross-validation. Results can be output in table and /or chart form with node content, label and color options. Pruning can be better conducted by specifying minimum risk / impurity reduction instead of at users' discretion. SPSS 14.0 provides very user-friendly platform with concrete and concise option tabs. It outputs misclassification results directly.

S-Plus 6.1 provides easy tree pruning by its simple programming codes. Model assessment graphics are based on misclassification or deviance results which can be read directly. S-Plus 6.1 provides an option for specifying validation data simultaneously to running training data. It is less plausible in terms of option settings and flexibilities.

In summary, consider both misclassification results of our data set and overall

functions in CART modeling, we recommend SPSS 14.0 or SAS EM 5.0 the first, depending on the needs and goals of particular studies. S-Plus is the least preferable.

4 DISCUSSION

4.1 Key Fndings

4.1.1 Clinical Fndings

The prevalence estimate of TE and Infection in children with ALL during the period between from 1995 and 2005 is 15.07% (9.27%, 20.87%) and 57.24% (49.19%, 65.29%).

TE and Infection are clinically associated with each other. Age category, CVL functionality, Site of CVL insertion, and Infection are four clinically important predictors of TE. ANC category, Age category, CVL functionality, and TE are four clinically important predictors of Infection.

4.1.2 Statistical Findings

In exploratory sense, TE and Infection are not statistically significantly associated with each other. No predictors are significant predictors of TE. CVL functionality and ANC category are significant predictors of Infection.

With and without multiple imputation technique incorporated, logistic regressions, Fisher's exact test and Bayesian analysis arrived at consistent results in screening significant predictors of TE and of Infection.

With and without multiple imputation technique incorporated, SAS EM 5.0, SPSS 14.0 and S-Plus 6.1, taking into account various tree growth methods and node splitting rules, resulted in consistent misclassification results in predicting TE. However, in predicting Infection, the averaged misclassification rates with multiple imputed data sets were lower than those resulted from the original data set. Compared to split sampling validation, 10-fold cross-validation with the multiple imputed data sets resulted in closer-to-original misclassifications. Misclassifications resulted by SPSS CHAID method were consistent to those resulted by SAS EM Chi-Square methods. This implies that in our study, Kass adjustment for multiple testing in CHAID did not make much difference with

Chi-Square testing. SPSS CRT with Gini or Twoing gave consistently lower misclassifications than all other techniques. SPSS QUEST algorithm resulted in considerably higher misclassifications than all other methods. This was partly because QUEST need to transform values of the predictors which were all categorical to monotonic numbers for modeling.

SAS EM 5.0 and SPSS 14.0 are more feasible and have more plausible options than S-Plus 6.1 in doing CART modeling. Their misclassification rates are more stable and accountable than those assessed by S-Plus 6.1.

4.2 Comparison of the Results with Those from Similar Studies

4.2.1 Clinical Results

We estimated the prevalence of TE in children with ALL from 1995 to 2005 to be 15.07%, with 95% confidence interval between 9.27% and 20.87%. The published estimates of prevalence of TE in children with ALL vary from the lowest 1.1% to the highest 36.7% with an overall average of 3.2%, most occurring within two to four years of study periods [1, 2, 3].

4.2.2 Statistical Results

Table 4.1 demonstrates a comparison of this study and other five published studies. Two studies [3, 32] identified CVL insertion technique and site of CVL insertion to be significant predictors of TE. The above studies screened all patients for asymptomatic TE whereas we only included clinically evident TE. One study [33] identified type of CVL to be a significant predictor of TE. Our study did not find any significant predictor of TE. Its results, however, might be impacted by the exclusion of Type of CVL and Type of Asparaginase at the beginning of the study.

As for association of TE with Infection, our study agrees with one study [34] in that there is no significant association between thrombophilia or infection and

development of TE. Most of these studies has the problem of small sample size.

Infection was studied as another outcome besides TE in our study. Our study is the only one which conducted CART modeling in parallel to traditional studies and compared results with theirs. In addition, from therapeutic point of view, CART modeling was conducted a second time to generate clinical importance ordered decision trees.

The robustness of results under classical logistic regression and Bayesian analysis when using non-informative priors in this study agrees with previous studies in that the two approaches lead to similar results [35-38].

4.3 Limitations of the Study

Attention need to be paid to some limitations of this study.

4.3.1 Small Sample Size

Sample size of 150 limited the precision of our analysis and resulted in wide confidence intervals. Some potential significant factors would not have been excluded from analysis if we had had larger sample. Small sample size also led to issues in CART modeling performance assessment, such as misclassification evaluation.

4.3.2 Missing Values

We conducted multiple imputation on categorical variables based on a compromised multi-normality assumption, which was not justifiable theoretically. Systematic bias existed along with the "filling-in" of missing data, though we used non-rounded imputation strategy in comparison.

4.3.3 Imbalance of Data

We considered Type of CVL and Type of Asparaginase for inclusion as predictors in the analysis. But we did not have adequate information on these variables. There is some evidence [1] to suggest that they may also be important variables to consider.

4.3.4 Time Point

The data used in our study covered a long period of 16 years. Time of Treatment might be a potential predictor for our outcome because treatment progressed as time passed. However, we could not include this into our analysis due to the unavailability of such information.

4.3.5 CART Modeling

With such a small sample size data, to estimate reliable misclassifications comparable among the three CART modeling packages, we would, on one hand include as many records as possible in estimating, on the other hand, get as little bias as possible with such internal (resubstitution) estimation. As a balanced approach, V-fold cross validation was the preferable assessment technique for our case. However, this option is provided with SPSS 14.0 only.

First, with the three packages, we separated data into training and validation / testing groups and then chose to assess misclassifications based on training and validation parts. The misclassification estimates were comparable among the packages, however with unavoidable downward bias. Last, we used 10-fold cross-validation with intention to reach less-biased misclassification estimates.

4.4 Implications of the Findings

4.4.1 Clinically: Hypothesis Generation

From this study, we hypothesize that

Age category, Site of CVL, CVL functionality, and Infection are clinically important to TE. ANC category, Age category, CVL functionality, and TE are clinically important to Infection. Among them, CVL functionality and ANC category are significant to Infection.

Further studies need big enough in sample size so that Type of CVL and Type of Asparaginase can be included in the analysis and their significance to TE and Infection

can be tested.

CART modelings for TE and for Infection by clinical importance order provide helpful insights into setting therapeutic or preventive protocols for each particular patient.

4.4.2 Statistically: Modeling

CART modeling can be a valuable tool in one's arsenal of data analysis tools. It provides valid and helpful supplement to the traditional approaches. However, CART models need to be used in conjunction with subject matter experts instead of being used in isolation.

Our software evaluations were a look at the potential each products offers in decision tree methodology. Accuracy, parsimony, non-trivial, feasibility, transparency and interpretability were the evaluation bases. However, the strengths and weaknesses of each package should be weighted according to the audience (statistician) and the problem domain (medicine, industry, etc.)

The comparisons of misclassifications by the three software packages were estimated with our project data only. General conclusion about accuracy need to be further studied with various patterns of data sets.

5 CONCLUSIONS

The prevalence estimate of TE in children with ALL from 1995 to 2005 is 15.07% (9.27%, 20.87%). The prevalence estimate of Infection in children with ALL from 1995 to 2005 is 57.24% (49.19%, 65.29%). In ALL children treated with CVL, ANC category and CVL functionality were identified as two significant factors leading to infection. No factors were identified to be significant to the incurrence of TE.

In terms of exploring significant factors of TE and of infection, simple logistic regression, logistic regression with rounding and non-rounding incorporated MI, Fisher's exact test, Bayesian analysis led to consistent results. The estimates of odds ratios by the two outcomes, as well as the estimates of CART modeling misclassification errors are generally consistent, with and without integrating multiple imputation approaches.

CART modeling is a valuable exploratory tool supplementary to traditional parametric analyses in exploring outcome predictors through generating intuitive and interpretable decision trees. Its value can be better exploited when being collaborative used by statistician and subject matter experts, which is essential in interactively growing and pruning the trees. Based on our data, SAS EM 5.0, SPSS 14.0 and S-Plus 6.1 led to consistent CART modeling accuracy. Although the common splitting sampling validation techniques we chose with the three packages introduced biases in estimating misclassification, the results were generally comparable to each other. SAS EM 5.0 and SPSS 14.0 are more plausible than S-Plus for their advantages in feasibility, transparency and interpretability. However, the strengths and weaknesses of each package should be weighted according to the users and the problem natures.

The limitations of this exploratory study such as small sample size, missing values, unbalance between data categories, ignorance of time point, lack of cross validation techniques of some packages in evaluating CART modeling performance led biases to our results. Large data set with few missing values set is critical to arrive at more accurate results. Meta analysis should be a worthy approach to further conduct on similar studies in this field selected by certain criteria given the low event rate of TE.

CART modeling provides visual prediction methods which can be easily applied in a clinical setting. But it does not provide estimates of measures of association to assess the strength of the association. Classical and Bayesian logistic models provide estimates of odds ratios (95% confidence intervals) and associated p-values. These results can also be used for developing prediction equations or rules for use in a clinical setting.

6 BIBLIOGRAPHY

[1] Athale UH, Chan AKC. "Thrombosis in Children with Acute Lymphoblastic Leukemia Part I: Epidemiology of Thrombosis in Children with Acute Lymphoblastic Leukemia." *Thrombosis Research* 2003; 111(3): 125-131.

[2] Mitchell LG, Sutor AH, Andrew M. "Hemostasis in Childhood Acute Lymphoblastic Leukemia: Coagulopathy Induced by Disease and Treatment." *Seminars in Thrombosis and Hemostasis* 1995; 21(4): 390-401.

[3] Mitchell LG, et al. "A Prospective Cohort Study Determining the Prevalence of Thrombotic Events in Children with Acute Lymphoblastic Leukemia and a Central Venous Line Who are Treated with L-Asparaginase: Results of the Prophylactic Antithrombin Replacement in Kids with Acute Lymphoblastic Leukemia Treated with Asparaginase (PARKAA) Study." *Cancer* 2003; 97(2): 508-16.

[4] Andrew M, et al. "Maturation of the Hemostatic System during Childhood." *Blood* 1992; 80: 1998-2005.

[5] Nowak-Gottl U, Kosch A, Schlegel N. "Thromboembolism in Newborns, Infants, and Children." *Journal of Thrombosis and Haemostasis* 2001; 86: 464-74.

[6] Monagle P, et al. "Antithrombotic Therapy in Children." *Chest* 2001; 119: 344-370S.

[7] Schafer JL. *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall/CRC, 1997.

[8] Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc., 1987.

[9] Rubin DB. "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association* 1996; 91: 473–489.

[10] Schafer JL. "Multiple Imputation: a Primer." *Statistical Methods in Medical Research* 1999; 8(1): 3-15.

[11] *Multiple Imputation Online.* < http://www.multiple-imputation.com/ >. Oct.13th, 2007

[12] Horton NJ, Lipsitz SR, Parzen M. "A Potential for Bias When Rounding in Multiple Imputation." *The American Statistician* November 2003.

[13] Ake CF. "Rounding After Multiple Imputation with Non-binary Categorical Covariates." *SUGI* 25: Paper 112-30.

[14] Yuan Yang C. "Multiple Imputation for Missing Data: Concepts and New Development." *SUGI* 30: Paper 267-25.

[15] SAS Institute. *SAS OnlineDoc™. Version 8. Chapter 9, The MI Procedure.*

[16] SAS Institute. *SAS OnlineDoc™: Version 8. Chapter 10, The MYANALYZE Procedure.*

[17] Berger JO. Statistical Decision Theory and Bayesian Analysis 2nd ed. New York: Springer-Verlag, 1985.

[18] Fredman L. "Bayesian Statistical Methods." *BMJ* 1996; 313: 569-570.

[19] Spiegelhalter DJ, et al. "Bayesian Methods in Health Technology Assessment: a Review." *Health Technology Assessment* 2000; 4(38).

[20] Landrum MB, Normand S-L. "Applying Bayesian Ideas to the Development of Medical Guidelines." *Statistics in Medicine* 1999; 18(2): 117 – 137.

[21] Sung L, et al. "Seven items Were Identified for Inclusion When Reporting a Bayesian Analysis of a Clinical Study." *Journal of Clinical Epidemiology* 2005; 58(3): 261-268.

[22] Robert CP, Casella G. *Monte Carlo Statistical Methods.* New York: Springer, 1999.

[23] *The Bugs Project.* < http://www.mrc-bsu.cam.ac.uk/bugs >. Jan 10th, 2008

[24] Breiman Leo, et al. *Classification and Regression Trees.* New York: Chapman and Hall/CRC, 1984.

[25] Terri Moore, Carole Jesse, Richard Kittler. "An Overview and Evaluation of Decision Tree Methodology." *ASA Quality and Productivity Conference* 2001.

[26] *Canadian Institutes of Health Research.* < http://www.cugr.ca/e/193.html >.

Dec 10th, 2007.

[27] Rubin DB. "Inference and Missing Data." *Biometrika* 1976; 63: 581-592.

[28] Shrive FM, et al. "Dealing with Missing Data in a Multi-Question Depression Scale: a Comparison of Imputation Methods." *BMC Medical Research Methodology* 2006; 6:57.

[29] Kass G.V. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics* 1980; 29 (2): 119-127.

[30] Loh W-Y, Shih Y-S. "Split Selection Methods for Classification Trees." *Statistica Sinica.* 1997 (7): 815-840.

[31] Lewis RJ. "Introduction to CART Analysis." *Annual Meeting of the Society for Academic Emergency Medicine* 2000. San Francisco.

[32] PARKAA Study. *Thromb Haemost.* 2002; 87(4): 593-8.

[33] McLean TW, et al. "Central venous lines in children with lesser risk acute lymphoblastic leukemia: optimal type and timing of placement." Journal of Clinical Oncology 2005; 23: 3024-3029.

[34] Rudd et al. "Prevalence of thrombophilia and central venous catheter-associated neck vein thrombosis in 41 children with cancer – a prospective study." *Med Pediatric Oncology* 2002; 38: 405-10

[35] Berger JO. "An Overview of Robust Bayesian Analysis." *Test* 1994; 3: 5-58.

[36] Berger JO. "Robust Bayesian Analysis: Sensitivity too the Prior." *Journal of Statistical Planning and Inference* 1990; 25: 303-28.

[37] Lavine M. "Sensitivity in Bayesian Statistics: the Prior and the Likelihood." *Journal of the American Statistical Association* 1991; 86: 396-99.

[38] Kass RE, et al. Approximate methods for Assessing Influence and Sensitivity in Bayesian Analysis. Biometrika 1989: 76: 663-74.

[39] *SEER Cancer Statistics Review, 1973-1992.* US Department of Health and Human Services, NIH Publication No. 96-2789 p.463

Table 2.1 Terminologies and Abbreviations

Abbreviation	Full Terminology	Explanation & Remarks
ALL	acute lymphoblasic leukymia	
Age_cate	age category	binary measurement
ANC	absolute neutrophil count	continuous measurement
ANC_cate	absolute neutrophil count category	binary measurement
Bodyside	site of CVL insertion	
Chemophase	phase of chemotherapy at CVL insertion	
CNS	central nervous system	
CVL	central venous catheter	
CART / CRT	classification and regression trees	developed by Brieman, Friedman, Olshen, and Stone in 1984
CHAID	chi-square automated interaction detection	developed by Kass (Applied Statistics, 1980)
DFCI	Dana-Farber Cancer Institute ALL Consortium	
DVT	deep venous thrombosis	
E.Coli	escherichiae coli	asparaginase type
Entropy		impurity reduction algorithm
Gini		impurity reduction algorithm
HR	high risk of TE	definition in Section 3
MAR	missing at random	
MCMC	Markov chain Monte Carlo simulation	
PE	pulmonary embolism	type of TE
PEG	polyethelyne glycosalated	asparaginase type
PICC	peripherally inserted central catheter	
QUEST	quick, unbiased, efficient, statistical tree	developed by Loh and Shih (Statistica Sinica, 1997)
SR	standard risk of TE	definition in Section 3
TE	thromboembolism	
Twoing		
WBC	white Blood Cell	

Table 2.2 Summary of Tree Growth & Node Splitting Rules

Feature	CHAID	CART	QUEST
Developer	Kass (Applied Statistics, 1980)	Brieman, Friedman, Olshen, and Stone (1984)	Loh and Shih (Statistica Sinica, 1997)
2-class / multi-class	both	both	2-class
Variable selection Method	X^2 test with Bonferroni adjustment	criteria of node splitting/remerging - Gini - Entropy - Twoing	F and X^2 tests with Bonferroni adjustment
Split point selection	splitting and remerging	exhaustive search	quadratic discriminant analysis (QDA))or exhaustive search
Predictor variable type	free - categorical - ordinal categorical with one exceptional category (e.g. missing value) Monotonic - numerical discretized to ordinal categorical	numerical Ordinal Categorical	monotonic - numerical - categorical transformed to numerical
Uni- / Multi- splits	univariate only	both	both
Pruning	top-down	bottom-up	top-down
Cost-complexity pruning (Cross validation)		yes	
Missing values		surrogate split	mean / mode imputation
Split of Entropy minimizes ($p(j/t)$: probability of node t classified to j)		$-\sum_j p(j \mid t) \times \log(p(j \mid t))$	
Split of Gini minimizes ($p(j/t)$: probability of node t classified to j)		$\sum_{i \neq j} p(i \mid t) \times p(j \mid t)$	
Split of Twoing maximizes		$\frac{P_L \times P_R}{4}\left[\sum_j \lvert p(j \mid t_L) - p(j \mid t_R)\rvert\right]^2$	

Table 2.3 Patient Demographics by TE

Variables	Statistics	Presence of TE	Absence of TE	Valid Number
Age (in years)	mean (sd)	7.27 (5.22)	6.27 (4.20)	146
Gender				
- female	count (%)	11(17.74)	51(82.26)	62
- male		11(13.10)	73(86.90)	84
ALL-risk				
- SR	count (%)	12 (14.12)	73 (85.88)	85
- HR		10 (16.67)	50 (83.33)	60
- missing		0	1	
Chemotherapy Phase at CVL Insertion				
- prior to start of chemo	count (%)	14 (19.18)	59 (80.82)	73
- induction		7 (10.61)	59 (89.39)	66
- CNS and intensification		1 (33.33)	2 (66.67)	3
- consolidation		0 (0)	2 (100)	2
- maintenance		0 (0)	1 (100)	1
- CNS prophylaxis & consolidation		0 (0)	1 (100)	1
- missing		0	0	
CVL Insertion Technique				
- percutaneous	count(%)	14 (14.29)	84 (85.71)	98
- open		8 (18.60)	35 (81.40)	43
- missing		0	5	
CVL Functionality				
- dysfunction		7 (20.59)	27 (79.41)	34
- function		14 (12.96)	94 (87.03)	108
- missing		1	3	
Type of Asparaginase				
- E. Coli	count (%)	15 (15.79)	80 (84.21)	95
- Erwinia		0 (0)	23 (100)	23
- PEG		0	0	0
- others (switches)		0	0	0
- missing		7	21	
Type of CVL				
- portacach	count (%)	19 (13.57)	121 (86.43)	140
- hichman		3 (60)	2 (40)	5
- PICC		0 (0)	1 (100)	1
Site of Insertion				
- left	count (%)	3 (8.11)	34 (91.89)	37
- right		19 (17.92)	87 (82.08)	106
- missing		0	3	

Table 2.4 Patient Demographics by Infection

Variables	Statistics	Presence of Infection	Absence of Infection	Valid Number
Age (in years)	mean (sd)	6.96 (4.79)	5.68 (3.65)	145
Gender				
- female	count (%)	36 (58.06)	26 (41.94)	62
- male		47 (56.63)	36 (43.37)	83
- missing		0	0	
ALL-risk				
- SR	count (%)	46 (55.42%)	37 (44.58%)	83
- HR		36 (59.02%)	25 (40.98%)	61
- missing		1 (100%)	0 (0%)	
Chemotherapy Phase at CVL Insertion				
- prior to start of chemo	count (%)	41 (57.75)	30 (42.25)	71
- induction		40 (59.70)	27 (40.30)	67
- CNS and intensification		0 (0)	3 (100)	3
- consolidation		1 (50)	1 (50)	2
- maintenance		0 (0)	1 (100)	1
- CNS prophylaxis & consolidation		1 (100)	0 (0)	1
- missing		0	0	
CVL Insertion technique				
- percutaneous	count (%)	55 (57.29)	41 (42.71)	96
- open		25 (56.82)	19 (43.18)	44
- missing		3	2	
CVL functionality				
- dysfunction	Count (%)	26 (74.29)	9 (25.71)	35
- function		54 (51.43)	51 (48.57)	105
- missing		3	2	
Type of CVL				
- portacach	count (%)	77 (55.40)	62 (44.60)	139
- hichman		5(100)	0 (0)	5
- PICC		1 (100)	0 (0)	1
- missing		0	0	
ANC at CVL Insertion	mean (sd)	1.97 (3.63)	1.79 (2.60)	136
- missing	count	4	5	
Site of Insertion				
- left	count (%)	19 (52.78)	17 (47.22)	36
- right		62 (59.05)	43 (40.95)	105
- missing		1	2	

Table 2.5 Predictor Summary and Coding for TE

Variable (full name)	Values	Number	Percent	Coding
TE (thromboembolism)	present	21	14.66%	1
	absent	124	2.67%	0
	unknown (missing)	4	82.67%	9999
infection (infection)	positive	83	55.33%	1
	negative	62	41.33%	0
	unknown	5	3.33%	9999
age_cont (continuous age) age_cate (categorical age)	continuous value	150	100.00%	
	age<=10	118	78.67%	1
	10< age <=18	32	21.33%	0
gender (gender)	female	64	42.67%	1
	male	86	57.33%	0
risk (ALL-risk)	HR	61	40.67%	1
	SR	87	58.00%	0
	unknown (missing)	2	1.33%	9999
chemophase (phase of chemotherapy at CVL Insertion)	before chemo	73	48.67%	1
	induction	68	45.33%	2
	CNS & intensification	3	2.00%	0
	consolidation	2	1.33%	
	maintenance	1	0.67%	
	relapse	0	0.00%	
	CNS prophylaxis & consolidation	1	0.67%	
	unknown (missing)	2	1.33%	9999
insertion (CVL insertion technique)	percutaneous	99	66.00%	1
	open	44	29.33%	0
	unknown (missing)	7	4.67%	9999
dysfunCVL (CVL functionality)	yes	35	23.33%	1
	no	108	72.00%	0
	unknown	7	4.67%	9999
asprgtype (type of asparaginase)	E. Coli	96	64.00%	1
	Erwinia	23	15.33%	0
	PEG	0	0.00%	
	unknown (missing)	31	20.67%	9999
typeCVL (type of CVL)	portacath	142	94.67%	1
	hickman	5	3.33%	
	PICC	1	0.67%	0
	unknown (missing)	2	1.33%	100
bodyside (site of CVL insertion)	left	37	24.67%	1
	right	108	72.00%	0
	unknown (missing)	5	3.33%	9999

Table 2.6 Predictor Summary and Coding for Infection

Variable (full name)	Values	Number	Percent	Coding
TE (thromembolism)	present	21	14.66%	1
	absent	124	2.67%	0
	unknown (missing)	4	82.67%	9999
infection (infection)	positive	83	55.33%	1
	negative	62	41.33%	0
	unknown	5	3.33%	9999
age_cont (continuous age) age_cate (categorical age)	continuous value	150	100.00%	
	age<=10	118	78.67%	1
	10 < age <=18	32	21.33%	0
gender (gender)	female	64	42.67%	1
	male	86	57.33%	0
risk (ALL_risk)	HR	61	40.67%	1
	SR	87	58.00%	0
	unknown (missing)	2	1.33%	9999
chemophase (phase of chemotherapy at CVL Insertion)	before chemo	73	48.67%	1
	induction	68	45.33%	2
	CNS & intensification	3	2.00%	0
	consolidation	2	1.33%	
	maintenance	1	0.67%	
	relapse	0	0.00%	
	CNS prophylaxis & consolidation	1	0.67%	
	unknown (missing)	2	1.33%	9999
insertion (CVL insertion technique)	percutaneous	99	66.00%	1
	open	44	29.33%	0
	unknown (missing)	7	4.67%	9999
dysfunCVL (CVL functionality)	yes	35	23.33%	1
	no	108	72.00%	0
	unknown	7	4.67%	9999
typeCVL (type of CVL)	portacath	142	94.67%	1
	hickman	5	3.33%	
	PICC	1	0.67%	0
	unknown (missing)	2	1.33%	100
ANC_cont (ANC continuous) ANC_cate (ANC categorical)	continuous values	139	92.67%	
	unknown	11	7.33%	9999
	ANC < 0.5	57	38.00%	1
	ANC >= 0.5	82	54.67%	0
	unknown	11	7.33%	9999

Table 3.1 Odds Ratio Estimates by TE

Factor	Logistic Regression without MI	Fisher's Exact Test	Bayesian Analysis	Logistic Regression with MI (Rounding)	Logistic Regression with MI (Non-Rounding)
Infection (present vs absent)	1.727 (0.657, 4.538)	1.727 (0.607, 5.363)	1.802 (0.676, 5.068)	1.733 (0.660,4.555)	1.733 (0.660, 4.555)
Age in Year (per year increase)	1.051 (0.952, 1.159)		1.050 (0.950, 1.159)	1.048 (0.950,1.157)	1.048 (0.950, 1.157)
Age Category (1–10yr vs others)	0.541 (0.199, 1.469)	0.541 (0.184, 1.751)	0.555 (0.202, 1.542)	0.544 (0.201, 1.663)	0.544 (0.201, 1.474)
Gender (female vs male)	1.431 (0.577, 3.553)	1.431 (0.518, 70.103)	1.435 (0.559, 3.550)	1.447 (0.578,3.620)	1.447 (0.578, 3.620)
All-risk (HR vs SR)	1.217 (0.488, 3.033)	1.217 (0.434, 3.339)	1.205 (0.458, 3.065)	1.268 (0.507,3.172)	1.256 (0.504, 3.130)
Phase of Chemotherapy at CVL Insertion (prior to vs post)	1.424 (0.158, 2.795)	1.424 (0.152, 70.103)	1.449 (0.204,14.077)	1.405 (0.604,3.267)	1.405 (0.604, 3.267)
Phase of Chemotherapy at CVL Insertion (induction vs post)	0.712 (0.074, 6.804)	0.712 (0.068, 37.365)	0.727 (0.110, 7.248)	0.736 (0.301,1.804)	0.736 (0.301, 1.804)
Phase of Chemotherapy at CVL Insertion (induction vs prior to)	0.5 (0.188, 1.328)	0.5 (0.160, 1.445)	2.116 (0.805, 5.972)	0.736 (0.301, 1.804)	0.736 (0.301, 1.804)
CVL Insertion Technique (percutaneous vs open)	0.729 (0.281, 1.893)	0.729 (0.258, 2.200)	0.746 (0.286, 1.970)	0.714 (0.275,1.627)	0.697 (0.271, 1.795)
CVLFunctionality (dysfunction vs function)	1.741 (0.638, 4.747)	1.741 (0.536, 5.170)	1.703 (0.602, 4.627)	1.842 (0.654, 5.183)	1.908 (0.687, 5.296)
Site of CVL Insertion (left vs right)	0.404 (0.112, 1.454)	0.404 (0.072, 1.514)	0.352 (0.082, 1.200)	0.403 (0.114,1.418)	0.397 (0.111, 1.422)
Type of Asparaginase (E.Coli vs Erwinia)	>999.999 (2.001, ∞)				
Type of CVL (portacach vs others)	0.105 (0.016, 0.668)	0.105 (0.009, 1.001)			

Table 3.2 Odds Ratio Estimates by Infection

Factor	Logistic Regression without MI	Fisher's Exact Test	Bayesian Analysis	Logistic Regression with MI (Rounding)	Logistic Regression with MI (Non-Rounding)
TE (present vs absent)	1.727 (0.657, 4.538)	1.727 (0.607, 5.363)	1.778 (0.680, 4.884)	1.733 (0.660,4.555)	1.733 (0.660, 4.555)
Age in Year (per year increase)	1.074 (0.991, 1.163)		1.077 (0.994, 1.169)	1.074 (0.992,1.164)	1.074 (0.992, 1.164)
Age Category (1–10 yr vs others)	0.533 (0.232, 1.228)	0.533 (0.207, 1.308)	0.520 (0.217, 1.189)	0.535 (0.232,1.235)	0.535 (0.232, 1.235)
Gender (female vs male)	1.061 (0.545, 2.063)	1.061 (0.518, 2.181)	1.056 (0.543, 2.069)	1.092 (0.563,2.119)	1.092 (0.563, 2.119)
All-risk (HR vs SR)	1.158 (0.593, 2.261)	1.158 (0.563, 2.392)	1.158 (0.590, 2.265)	1.166 (0.600,2.266)	1.152 (0.590, 2.251)
Chemotherapy Phase At CVL Insertion (prior to vs post)	3.416 (0.620, 18.810)	3.417 (0.509, 37.591)	3.423 (0.651,20.352)	1.408 (0.729,2.722)	1.408 (0.729, 2.722)
Chemotherapy Phase At CVL Insertion (induction vs post)	3.703 (0.669, 20.489)	3.74 (0.547, 40.844)	3.722 (0.658, 23.017)	1.562 (0.806,3.024)	1.562 (0.806, 3.024)
Phase of Chemotherapy at CVL Insertion (induction vs prior to)	1.084 (0.550, 2.136)	1.084 (0.521, 2.259)	0.912 (0.458, 1.820)	1.562 (0.806, 3.024)	1.562 (0.806, 3.024)
CVL Insertion Technique (percutaneous vs open)	1.020 (0.496, 2.096)	1.020 (0.463, 2.223)	1.022 (0.496, 2.115)	0.978 (0.471, 2.032)	1.001 (0.483, 2.075)
CVL Functionality (dysfunction vs function)	**2.728 (1.167, 6.378) ***	**2.728 (1.100, 7.231)**	**2.815 (1.221,7.008) ***	**2.896 (1.228,6.828) ***	2.929 (1.235, 6.947)
ANC (per unit increase)	1.017 (0.913, 1.133)		1.021 (0.915, 1.148)	1.019 (0.914, 1.135)	1.023 (0.915, 1.145)
ANC Category (<0.5 vs >=0.5)*	**2.180 (1.063, 4.475)***	**2.181 (1.006, 4.802)**	**2.202 (1.064,4.513) ***	1.974 (0.943, 4.134)	1.974 (0.943, 4.131)
Site of CVL Insertion (left vs right)	0.775 (0.362, 1.660)	0.775 (0.338, 1.789)	0.758 (0.347, 1.626)	0.754 (0.338, 1.682)	0.732 (0.338, 1.587)
Type of CVL (portacach vs others)	<0.001 (0, 0.480)				

Table 3.3 Misclassification Comparison – Outcome TE

Software & Method		Misclassification Rates						
Software	Node Splitting Rules	Original Data Set	MI 1	MI 2	MI 3	MI 4	MI 5	MI Average
SAS EM 5.0 Split Sampling Training 90% Validation 5% Test 5% Max. depth: 3	Chi-Square	15.11%	15.38%	15.38%	15.38%	16.08%	16.08%	15.66%
	Entropy	15.11%	15.38%	15.38%	15.38%	16.08%	16.08%	15.66%
	Gini	15.11%	15.38%	15.38%	15.38%	16.08%	15.38%	15.52%
	Average	15.11%	15.38%	15.38%	15.38%	16.08%	15.85%	15.61%
SPSS 14.0 Split Sampling Training 90% Test 10% Max. depth: 3	CHAID (Pearson)	15.07%	15.33%	15.33%	15.33%	15.33%	15.33%	15.33%
	CHAID (Likelihood Ratio)	15.07%	15.33%	14.67%	15.33%	16.00%	15.33%	15.33%
	CRT (Gini)	14.38%	15.33%	15.33%	15.33%	16.00%	15.33%	15.46%
	CRT (Twoing)	15.07%	15.33%	14.67%	15.33%	16.00%	15.33%	15.33%
	QUEST	15.07%	15.33%	15.33%	15.33%	16.00%	15.33%	15.46%
	Sub-Average	14.93%	15.33%	15.07%	15.33%	15.87%	15.33%	15.38%
SPSS 14.0 10-fold validation Training 90% Test 10% Max. depth: 3	CHAID (Pearson)	15.07%	15.33%	14.67%	15.33%	15.33%	15.33%	15.20%
	CHAID (Likelihood Ratio)	15.07%	15.33%	14.67%	15.33%	15.33%	15.33%	15.20%
	CRT (Gini)	14.38%	15.33%	14.67%	15.33%	15.33%	15.33%	15.20%
	CRT (Twoing)	14.38%	15.33%	14.67%	15.33%	15.33%	15.33%	15.20%
	QUEST	15.07%	15.33%	15.33%	15.33%	16.00%	15.33%	15.46%
	Sub-Average	14.79%	15.33%	14.80%	15.33%	15.46%	15.33%	15.25%
S-PLUS 6.1		15.79%	15.33%	14.67%	15.33%	15.33%	15.33%	15.20%

CART / CRT : Classification and Regression Trees; QUEST: another type of decision tree
CHAID: Chi-Square Automated Interaction Detection MI: multiple imputation
Gini, Entropy and Towing: Impurity Reduction Algorithms

Table 3.4 Misclassification Comparison – Outcome Infection

Software & Method		Misclassification Rates						
Software	Node Splitting Rules	Original Data Set	MI 1	MI 2	MI 3	MI 4	MI 5	MI Average
EM 5.0 Split Sampling Training 90% Validation 5% Test 5% Max. depth: 5	Chi-Squre	42.75%	43.36%	42.66%	44.06%	34.97%	38.46%	40.70%
	Entropy	43.55%	43.36%	35.66%	35.66%	34.97%	41.26%	38.18%
	GINI	39.58%	35.66%	35.66%	35.66%	34.97%	41.26%	36.64%
	Average	41.96%	40.79%	37.99%	38.46%	34.97%	40.33%	38.51%
SPSS 14.0 Split Sampling Training 90% Test 10% Max. depth: 5	CHAID (Pearson)	45.21%	34.67%	31.33%	45.50%	35.33%	32.67%	35.90%
	CHAID (Likelihood Ratio)	41.10%	33.33%	32.00%	40.00%	34.67%	38.00%	35.60%
	CRT (Gini)	31.51%	34.67%	30.00%	44.40%	31.33%	34.00%	34.88%
	CRT (Twoing)	32.88%	34.00%	31.33%	56.20%	35.33%	35.33%	38.44%
	QUEST	42.47%	43.33%	42.67%	44.00%	44.00%	41.33%	43.07%
	Sub-Average	38.63%	36.00%	33.47%	46.02%	36.13%	36.27%	37.58%
SPSS 14.0 10-fold validation Training 90% Test 10% Max. depth: 5	CHAID (Pearson)	31.51%	31.33%	28.67%	31.33%	31.33%	30.00%	30.53%
	CHAID (Likelihood Ratio)	31.51%	31.33%	28.67%	31.33%	31.33%	30.00%	30.53%
	CRT (Gini)	29.45%	33.33%	28.67%	33.33%	31.33%	31.33%	31.60%
	CRT (Twoing)	29.45%	33.33%	28.67%	33.33%	31.33%	31.33%	31.60%
	QUEST	42.47%	43.33%	42.67%	44.00%	44.00%	41.33%	43.07%
	Sub-Average	32.88%	34.53%	31.47%	34.66%	33.86%	32.80%	33.47%
S-PlusS 6.1		29.37%	32.00%	32.67%	32.00%	32.67%	32.67%	32.40%

CART / CRT : Classification and Regression Trees; QUEST: another type of decision tree
CHAID: Chi-Square Automated Interaction Detection MI: multiple imputation
Gini, Entropy and Towing: Impurity Reduction Algorithms

Table 3.5 Software Packages Comparison in CART Modeling

Feature		SAS EM 5.0	SPSS 14.0	S-PLUS 6.1
tree growth method & splitting rule	**Ci-Square**	Chi-Squre - Pearson / significance level	CHAID / Exhaustive CHAID significance level with - Pearson - Likelihood Ratio	- Pearson - Likelihood Ratio
	CART	Node Splitting Method - Gini - Entropy	Node Splitting Method - Gini - Twoing - Ordered Twoing	NA
	QUEST	NA	QUEST - significance level	NA
validation flexibility		- data needs to be partitioned to training, validation, testing groups - cross validation	- partition (training & validation) is optional not a must - cross validation	- no need to specify partition - cross validation
surrogate for missing		Applicable for - Chi-Square - CART with Entropy and Gini	Applicable for - CHAID and CRT -for QUEST, missing values are treated as floating category that is allowed to merge with other categories in tree nodes	yes
missing data		can be omit -can be treated as valid	- can be omit - can be treated as valid	can be omit can be treated as valid
statistics output		predicted value within-node yes / no proportion	misclassification (risk) rate summary table	misclassification errors Pearson residuals Deviance residuals number of terminal nodes
pruning		applicable for CART flexible pruning and recovery by clicking	applicable for CRT and QUEST by setting maximum difference of risks	"best=" command option for pruning
tabs for conducting CART modeling		Solutions→ Analysis → Enterprise Miner→Tools Input Data Source → Insight → Transform Varialbes → Data Set Attributes → Data Partition → Tree → Assessment → Report	Analyze→Classify→Tree Output→Validation→Criteria → Save	Statistics→ Tree → Tree→ Models Model→Result→Plot→Prune / Shrink→Predict

Table 3.5 Software Packages Comparison in CART Modeling (Continued)

Feature	SAS EM5.0	SPSS 14.0	S-Plus 6.1
Options	1. data partition methods: simple random / stratified / user defined. 2. training/validation/testing percentage setting 3. sampling process initiation seed 4. missing values replacement mechanism 5. significance level with Chi-Squre method 6. minimum obs in a leaf 7. obs required for a split search 8. maximum number of branches from a node 9. maximum depths of tree 10. splitting rules saved in each node 11. surrogate rules saved in each node 12. ignore or treating missing value as valid data. 13. model assessment measure: automatic; proportion misclassified; total leaf impurity (Gini index) 15. P-value adjustment 16. prior proabililty; cost / profit 17. Scores	1. maximum depths: automatic or custom 2. CHAID significance level for splitting / merge nodes 3. maximum number of iterations 4. minimum change in expected cell frequencies 5. saved variable: terminal nodes number, predictive value, predictive probability 6. P-value adjustment (Bonferroni) for CRT and QUEST 6. prior probability 7. cost / profit 8. scores 9. importance to model 10. forcing the first independent variable	1. min. number of obs before split 2. min. node size 3. min. node deviance 4. P-value adjustment (Bonferroni) 5. plot text label: response-variable / node size / node deviance 6. cost complexity pruning: cost complexity pruning parameter / size of returned tree 7. pruning / shrink method: cost complexity / optimal recursive shrinking / deviance / misclass 8. summary: summary description / full tree / print or save summary: misclassification error / Pearson residuals / deviance residuals 9. plot: branch size / branch label
Advantages	1. Easy comparison between logistic regression and CART 2. Easy pruning and recovery 3. Flexible nodes choice at user's discretion at any level 4. Convenient switch of data sources for tree modeling with saved project 5. Data set, attribute settings, partitions, tree options are connected as a project. CART modeling can be conducted with any of them adjusted and others unchanged. 6. Supports both automatic and interactive training of the model.	1. Data grouping is an option not a must. 2. Node content combines table and chart with label definition and color option. 3. Pruning is better conducted by specifying minimum risk / impurity reduction. 4. User friendly platform with concrete and concise option tabs 5. Direct misclassification results	1. Easy pruning 2. Direct summary of misclassification and deviance 3. Model assessment graphics based on misclassification or deviance 4. Option for specifying validation data simultaneously to running training data
	**	***	*

Table 4.1 Comparison of ALL Studies

Study	N	Risk Factors Identified
PARKAA Study *Thromb Haemost.* 2002; 87(4): 593-8.	66	CVL insertion technique, Site of CVL insertion
Mitchell LG, et al. "A prospective cohort study determining the prevalence of thrombotic events in children with acute lymphoblastic leukemia and a central venous line who are treated with L-asparaginase: results of the prophylactic antithrombin replacement in kids with acute lymphoblastic leukemia treated on asparaginase (PARKAA protocol)." *Cancer* 2003; 97(2): 508-16.	60	CVL insertion technique site of insertion
McLean TW, et al. "Central venous lines in children with lesser risk acute lymphoblastic leukemia: optimal type and timing of placement" POG (Pediatric Oncology Group) protocol 9201. *Journal of Clinical Oncology* 2005; 23: 3024-3029	362	Type of CVL
Rudd et al Prevalence of thrombophilia and central venous catheter-associated neck vein thrombosis in 41 children with cancer – a prospective study. *Med Pediatr Oncol* 2002; 38: 405-10	41	No correlation between thrombophilia or infection and development of TE
Our study by Logistic and Bayesian analyses	150	Significant risk factors for TE: none Infection: CVL functionality, ANC category
Our study by CART modeling	150	Misclassification Clinical importance ordered TE and infection decision trees Software comparison in doing CART modeling

Figure 1.1 Study Diagram

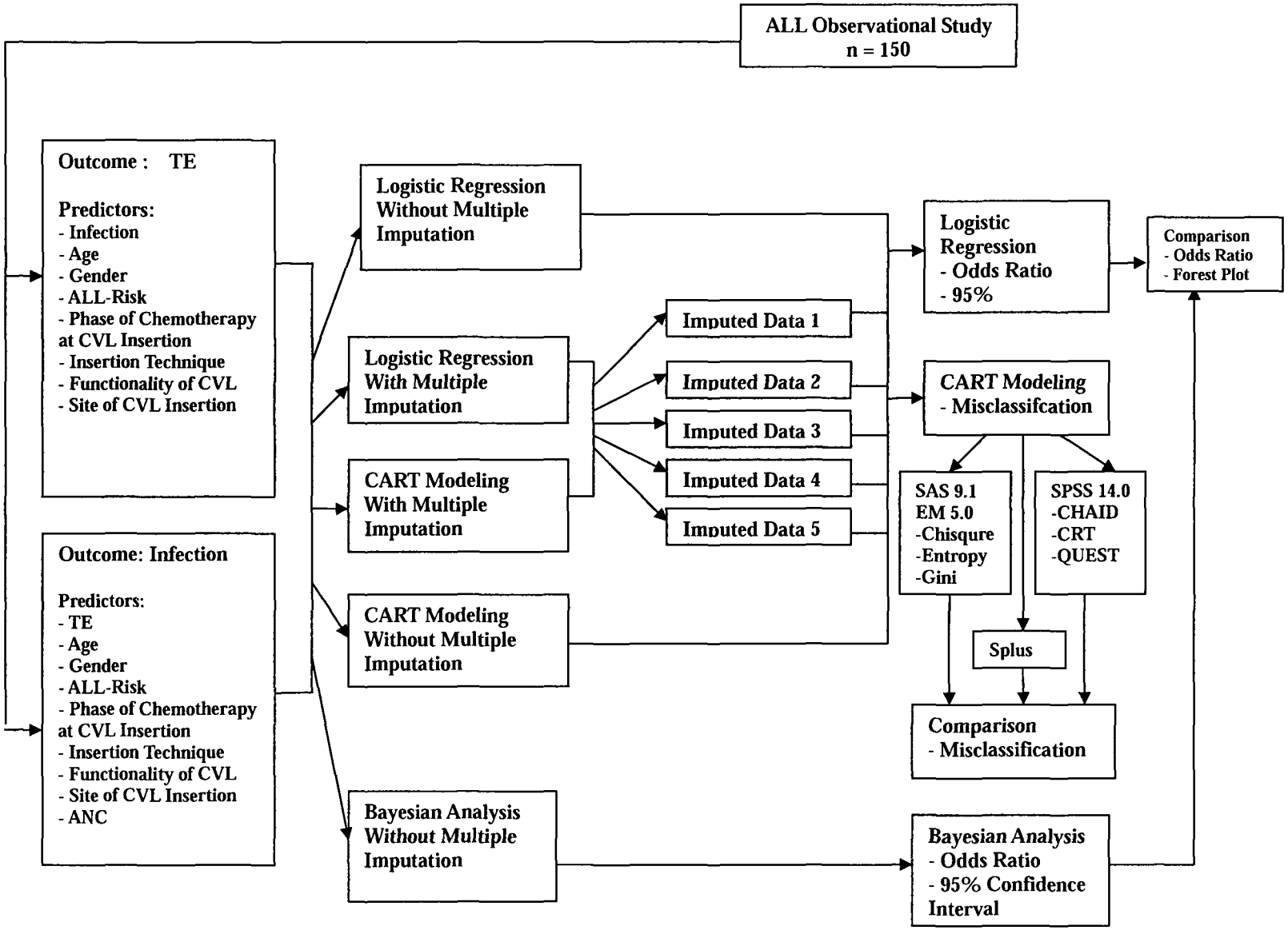
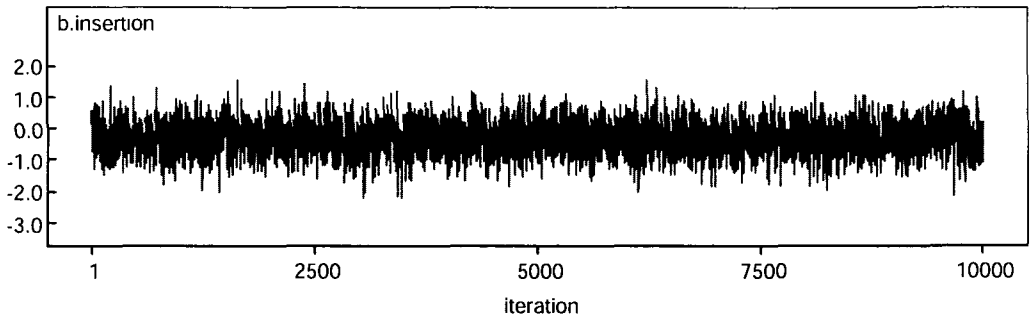
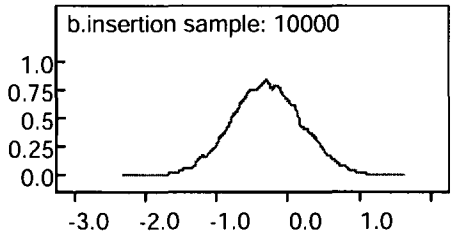
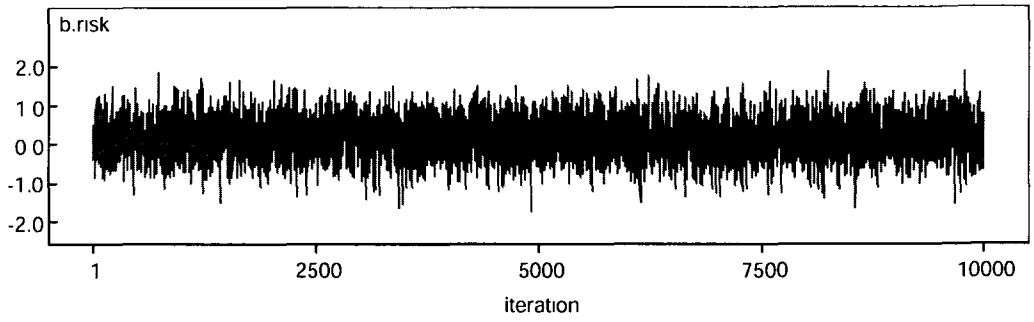


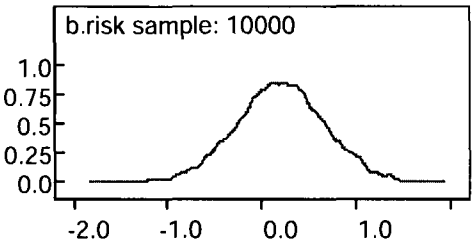
Figure 3.1 Bayesian Odds Ratio Examples by TE





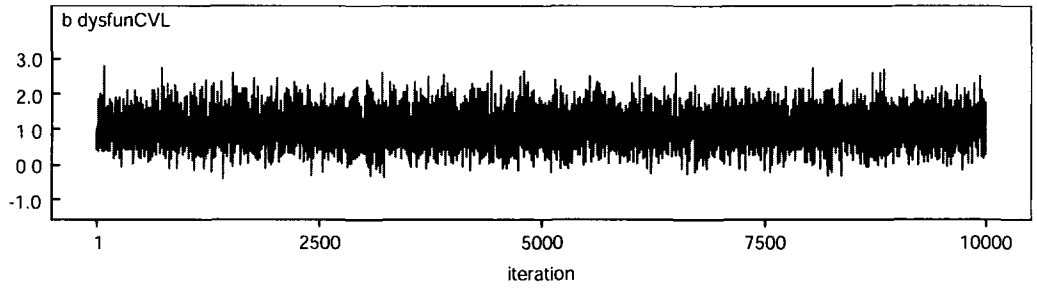
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b.insertion	-0.2937	0.4939	0.01024	-1.253	-0.2959	0.6779	1	10000

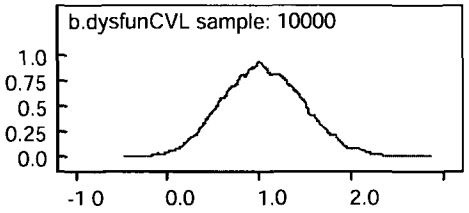




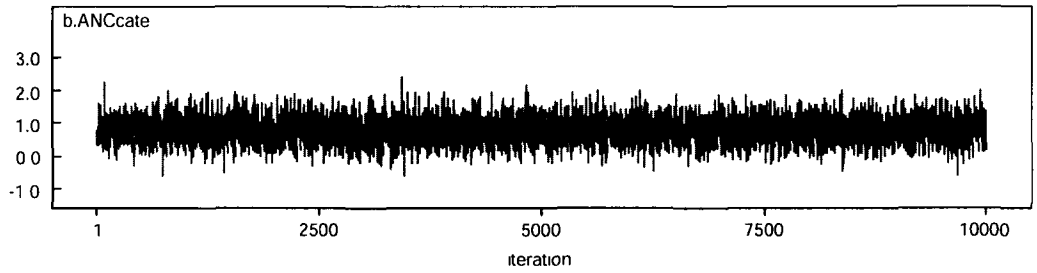
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b.risk	0.1861	0.4784	0.008184	-0.7814	0.1936	1.12	1	10000

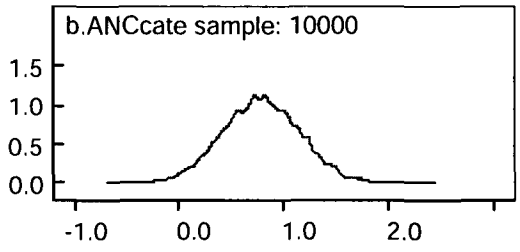
Figure 3.2 Bayesian Odds Ratio Examples by Infection





node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b.dysfunCVL	1.035	0.4444	0.004181	0.1994	1.02	1.947	1	10000



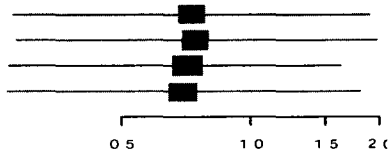


node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b.ANCcate	0.7892	0.3698	0.005206	0.06178	0.7903	1.507	1	10000

Figure 3.3 Forest Plot Example for Odds Ratios by TE

CVL Insertion by TE Odds Ratio (percutaneous to open)

Method	OR	95% Lower	95% Upper
Logistic w/o MI	0.729	0.281	1.893
Bayesian	0.746	0.286	1.97
Logistic MI&Rounding	0.714	0.275	1.627
Logistic MI w/o Rounding	0.697	0.271	1.795



At-Insertion Chemophase by TE Odds Ratio (induction to post chemotherapy)

Method	OR	95% Lower	95% Upper
Logistic w/o MI	0.712	0.074	6.804
Bayesian	0.727	0.11	7.248
Logistic MI&Rounding	0.736	0.301	1.804
Logistic MI w/o Rounding	0.736	0.301	1.804

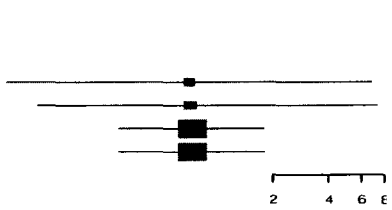
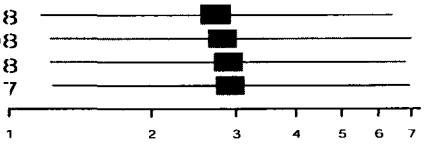


Figure 3.4 Forest Plot Example for Odds Ratios by Infection

CVL Functionality by Infection Odds Ratio (dysfunction to function)

Method	OR	95% Lower	95% Upper
Logistic w/o MI	2.728	1.167	6.378
Bayesian	2.815	1.221	7.008
Logistic MI&Rounding	2.896	1.228	6.828
Logistic MI w/o Rounding	2.929	1.235	6.947



ANC Category by Infection Odds Ratio (less than 0.5X10^9/L to others)

Method	OR	95% Lower	95% Upper
Logistic w/o MI	2.18	1.063	4.475
Bayesian	2.202	1.064	4.513
Logistic MI&Rounding	1.974	0.943	4.134
Logistic MI w/o Rounding	1.974	0.943	4.131

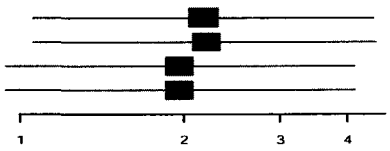
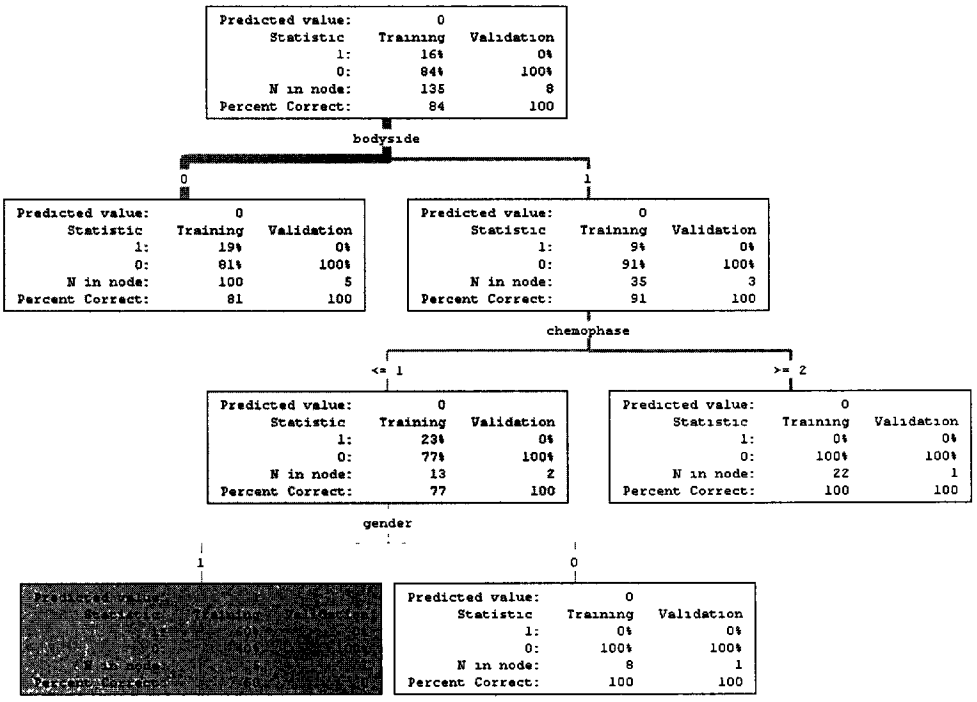


Figure 3.5 CART Modeling Tree Example for TE by Using SAS EM 5.0

Mutiple Imputed Data –TE (Chi Square)



leaf statistics				
training cases	validation cases	training correct percent	validation correct percent	misclassified cases
100	5	81	100	19
5	1	60	0	3
8	1	100	100	0
22	1	100	100	0
misclassification rate				**15.38%**

Figure 3.6 CART Modeling Example for TE by Using SPSS 14.0 Original Data

Model Summary

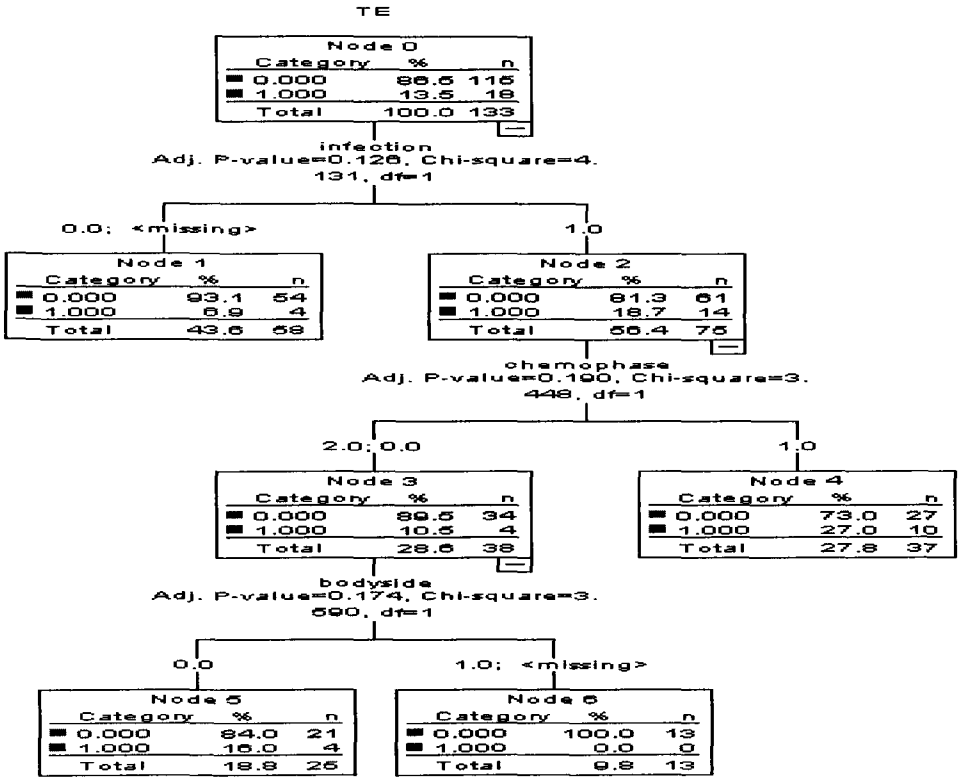
Specifications	Growing Method	CHAID (Likelihood Ratio)
	Dependent Variable	TE
	Independent Variables	agecate, gender, bodyside, risk, chemophase, insertion, dysfunCVL, infection
	Validation	Split Sample (Training 90%, Test 10%)
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	infection, chemophase, bodyside
	Number of Nodes	7
	Number of Terminal Nodes	4
	Depth	3

Risk

Sample	Estimate	Std. Error
Training	.135	.030
Test	.308	.128

Growing Method: CHAID

Dependent Variable: TE



Classification

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	115	0	100.0%
	1	18	0	.0%
	Overall Percentage	100.0%	.0%	86.5%
Test	0	9	0	100.0%
	1	4	0	.0%
	Overall Percentage	100.0%	.0%	69.2%

Growing Method: CHAID

Dependent Variable: TE

Misclassification

15.07%

Figure 3.7 CART Modeling Example for TE by Using S-plus 6.1

Original Data

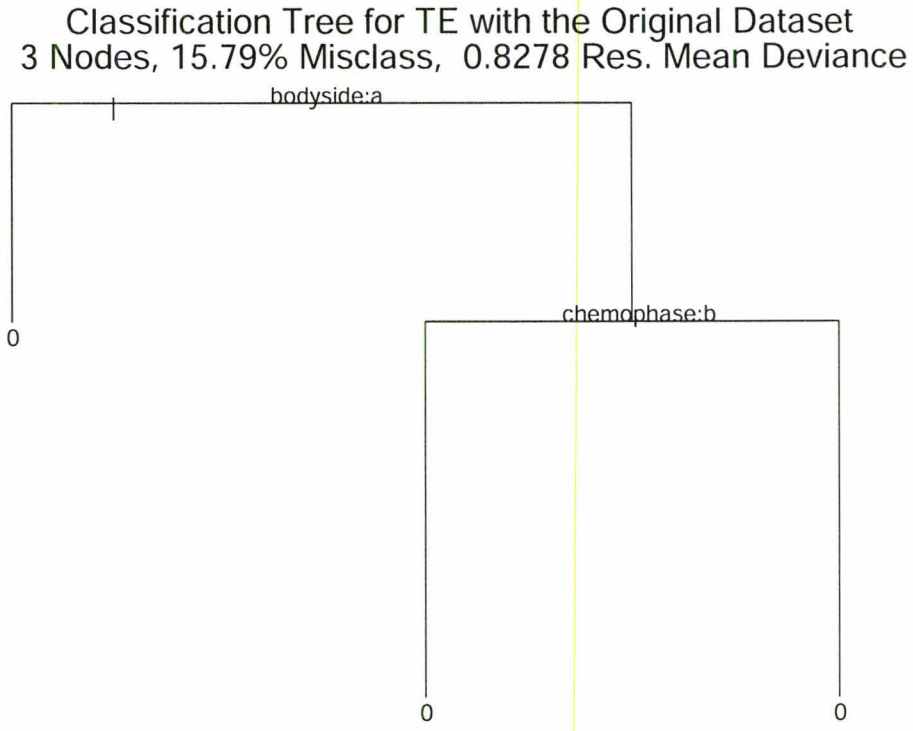
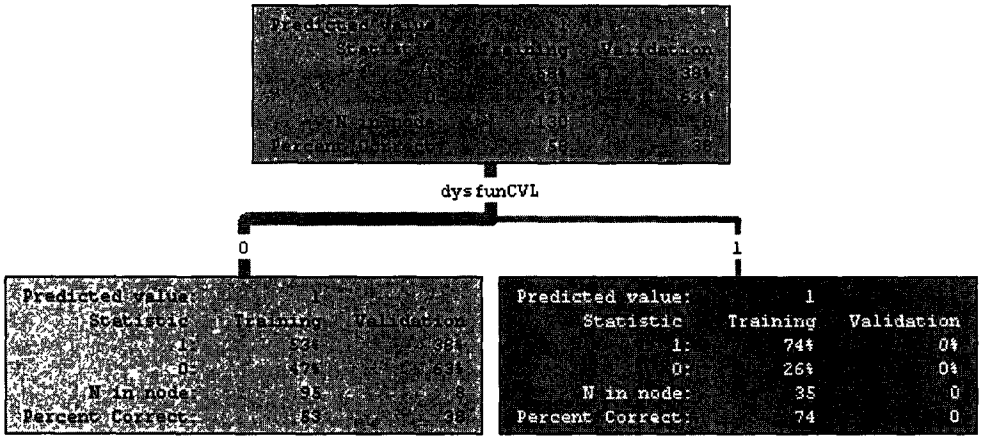


Figure 3.8 CART Modeling Example for Infection by SAS EM 5.0

Original data – Chi Square



leaf statistics				
training cases	validation cases	training correct percent	validation correct percent	misclassified cases
95	8	52.6316	37.5	50
35	0	74.2857	0	9
misclassification rate				**42.75%**

Figure 3.9 CART Modeling Example for Infection by Using SPSS 14.0

Model Summary

Specifications	Growing Method	CHAID (Pearson)
	Dependent Variable	infection
	Independent Variables	agecate, gender, bodyside, risk, chemophase, insertion, ANCcate, dysfunCVL, TE
	Validation	Cross Validation (Training 90%, Test 10%)
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	dysfunCVL, ANCcate, gender, risk, agecate
	Number of Nodes	17
	Number of Terminal Nodes	9
	Depth	4

Risk

Method	Estimate	Std. Error
Resubstitution	.317	.039
Cross-Validation	.531	.041

Growing Method: CHAID

Dependent Variable: infection

Classification

Observed	Predicted		
	0	1	Percent Correct
0	34	28	54.8%
1	18	65	78.3%
Overall Percentage	35.9%	64.1%	68.3%

Growing Method: CHAID

Dependent Variable: infection

Misclassification

31.51%

Figure 3.10 CART Modeling Example for Infection by Using S-plus 6.1
Original Data

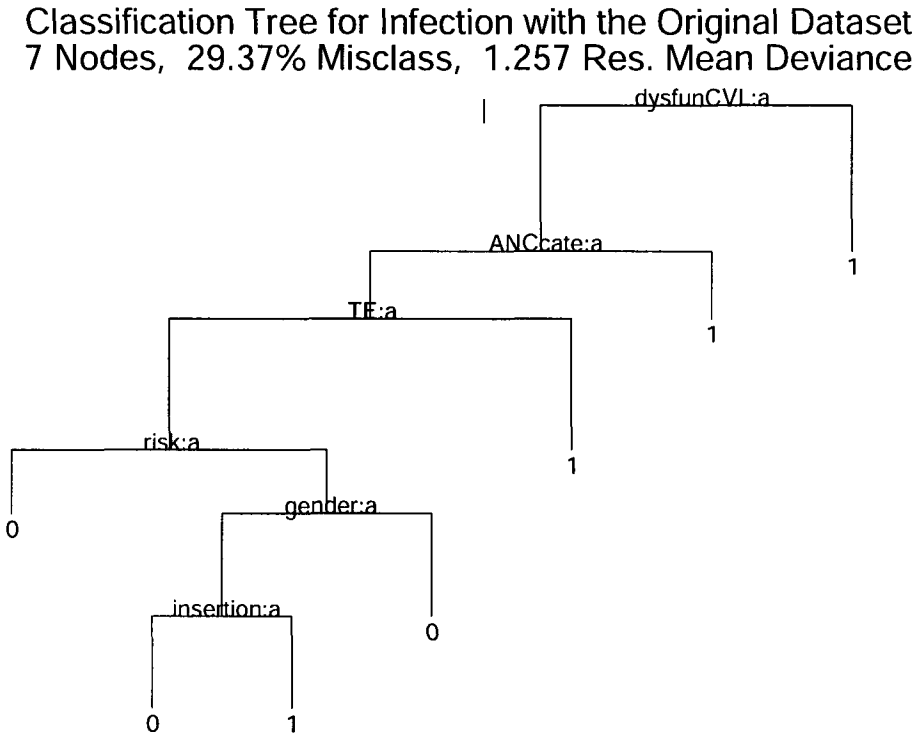
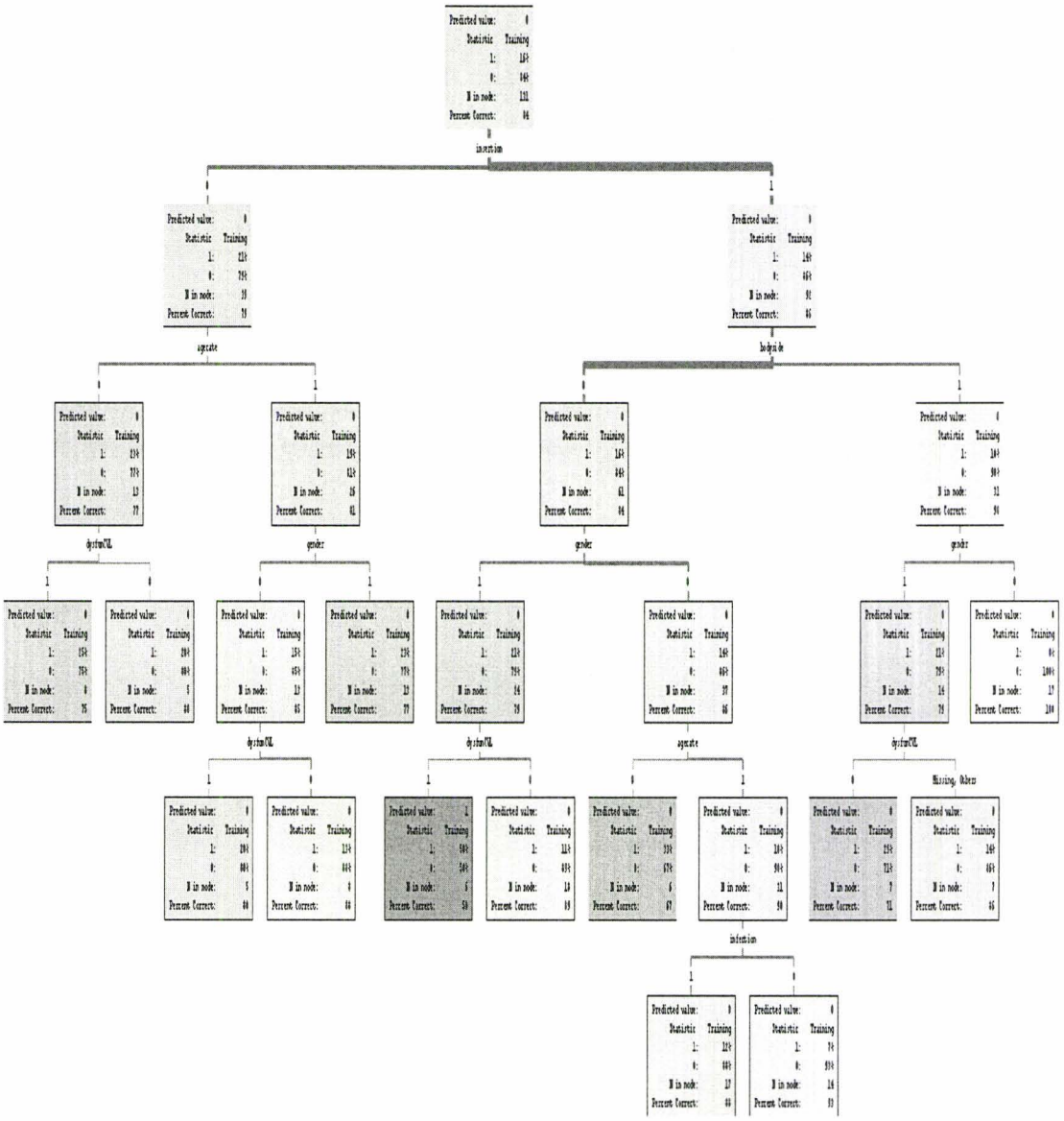


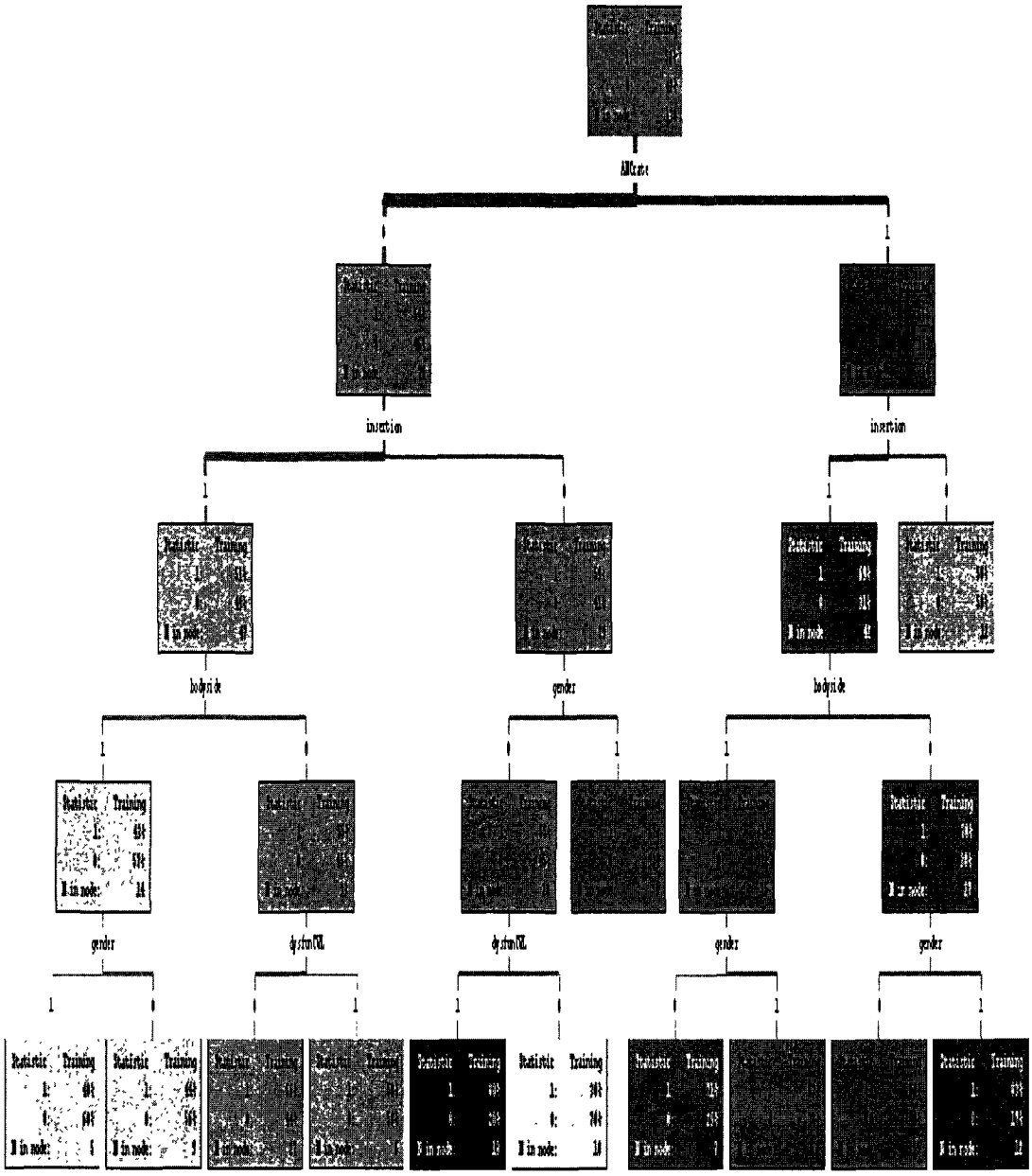
Figure 3.11 Clinical Importance Ordered Tree Example by SAS EM 5.0

Original Data – TE (Gini)
(Insertion– Site of Insertion– Gender & Age category – CVL functionality– Infection)



leaf statistics				
training cases	**validation cases**	**training correct percent**	**validation correct percent**	**misclassified cases**
8	0	75	0	2
5	1	80	100	1
13	2	76.9231	100	3
5	0	80	0	1
8	0	87.5	0	1
6	0	50	0	3
18	1	88.8889	100	2
6	1	66.6667	100	2
17	0	88.2353	0	2
14	1	92.8571	100	1
17	0	100	0	0
7	2	71.4286	100	2
7	0	85.7143	0	1
misclassification rate				**15.11%**

Original Data – Infection (Gini)
(ANC – Insertion – Site of Insertion – Gender – CVL functionality – TE)



leaf statistics				
training cases	**validation cases**	**training correct percent**	**validation correct percent**	**misclassified cases**
12	2	50	50	7
5	1	60	100	2
9	0	55.5556	0	4
27	3	55.5556	33.3333	14
6	0	50	0	3
9	1	66.6667	0	4
10	0	80	0	2
10	1	70	0	4
15	0	66.6667	0	5
27	0	70.3704	0	8
misclassification rate				**38.41%**

7 APPENDICES

7.1 R Code: Logistic Regressions

```
data<-read.table("data_all.txt",na.string="",sep="\t",header=T)
te.infection<-glm(data=data,TE~infection,family=binomial(link="logit"))
summary(te.infection)
exp(te.infection$coefficient[2])
exp(te.infection$coefficient[2]+c(-1,1)*1.96*0.4929)

te.agecont<-glm(data=data,TE~agecont,family=binomial(link="logit"))
summary(te.agecont)
exp(te.agecont$coefficient[2])
exp(te.agecont$coefficient[2]+c(-1,1)*1.96*0.05009)

te.agecate<-glm(data=data,TE~agecate,family=binomial(link="logit"))
summary(te.agecate)
exp(te.agecate$coefficient[2])
exp(te.agecate$coefficient[2]+c(-1,1)*1.96*0.5095)

te.gender<-glm(data=data,TE~gender,family=binomial(link="logit"))
summary(te.gender)
exp(te.gender$coefficient[2])
exp(te.gender$coefficient[2]+c(-1,1)*1.96*0.4638)

te.risk<-glm(data=data,TE~risk,family=binomial(link="logit"))
summary(te.risk)
exp(te.risk$coefficient[2])
exp(te.risk$coefficient[2]+c(-1,1)*1.96*0.4659)

te.chemophase<-glm(data=data,TE~factor(chemophase),family=binomial(link="logit"))
summary(te.chemophase)
exp(te.chemophase$coefficient[2])
exp(te.chemophase$coefficient[2]+c(-1,1)*1.96*1.1203)
exp(te.chemophase$coefficient[3])
exp(te.chemophase$coefficient[3]+c(-1,1)*1.96*1.1517)
data1<-data[data$chemophase!=0,]
te.chemophase12<-glm(data=data1,TE~factor(chemophase),family=binomial(link="logit"))
summary(te.chemophase12)
exp(te.chemophase12$coefficient[2])
exp(te.chemophase12$coefficient[2]+c(-1,1)*1.96*0.4982)

te.insertion<-glm(data=data,TE~insertion,family=binomial(link="logit"))
summary(te.insertion); exp(te.insertion$coefficient[2])
exp(te.chemophase$coefficient[2]+c(-1,1)*1.96*0.4867)

te.dysfunCVL<-glm(data=data,TE~dysfunCVL,family=binomial(link="logit"))
summary(te.dysfunCVL)
exp(te.dysfunCVL$coefficient[2])
exp(te.chemophase$coefficient[2]+c(-1,1)*1.96*0.5118)
```

```
te.bodyside<-glm(data=data,TE~bodyside,family=binomial(link="logit"))
summary(te.bodyside); exp(te.bodyside$coefficient[2])
exp(te.bodyside$coefficient[2]+c(-1,1)*1.96*0.6534)

infection.te<-glm(data=data, infection~TE, family=binomial(link="logit"))
summary(infection.te)
exp(infection.te$coefficient[2])
exp(infection.te$coefficient[2]+c(-1,1)*1.96*0.4929)

infection.agecont<-glm(data=data,infection~agecont,family=binomial(link="logit"))
summary(infection.agecont)
exp(infection.agecont$coefficient[2])
exp(infection.agecont$coefficient[2]+c(-1,1)*1.96*0.04083)

infection.agecate<-glm(data=data,infection~agecate,family=binomial(link="logit"))
summary(infection.agecate)
exp(infection.agecate$coefficient[2])
exp(infection.agecate$coefficient[2]+c(-1,1)*1.96*0.4255)

infection.gender<-glm(data=data,infection~gender,family=binomial(link="logit"))
summary(infection.gender)
exp(infection.gender$coefficient[2])
exp(infection.gender$coefficient[2]+c(-1,1)*1.96*0.33955)

infection.risk<-glm(data=data,infection~risk,family=binomial(link="logit"))
summary(infection.risk)
exp(infection.risk$coefficient[2])
exp(infection.risk$coefficient[2]+c(-1,1)*1.96*0.3414)

infection.chemophase<-
glm(data=data,infection~factor(chemophase),family=binomial(link="logit"))
summary(infection.chemophase)
exp(infection.chemophase$coefficient[2])
exp(infection.chemophase$coefficient[2]+c(-1,1)*1.96*0.8705)
exp(infection.chemophase$coefficient[3])
exp(infection.chemophase$coefficient[3]+c(-1,1)*1.96*0.8729)
data1<-data[data$chemophase!=0,]
infection.chemophase12<-
glm(data=data1,infection~factor(chemophase),family=binomial(link="logit"))
summary(infection.chemophase12)
exp(infection.chemophase12$coefficient[2])
exp(infection.chemophase12$coefficient[2]+c(-1,1)*1.96*0.34606)

infection.insertion<-glm(data=data,infection~insertion,family=binomial(link="logit"))
summary(infection.insertion)
exp(infection.insertion$coefficient[2])
exp(infection.insertion$coefficient[2]+c(-1,1)*1.96*0.3677)

infection.dysfunCVL<-glm(data=data,infection~dysfunCVL,family=binomial(link="logit"))
summary(infection.dysfunCVL)
exp(infection.dysfunCVL$coefficient[2])
exp(infection.dysfunCVL$coefficient[2]+c(-1,1)*1.96*0.43324)
```

```
infection.bodyside<-glm(data=data,infection~factor(bodyside),family=binomial(link="logit"))
summary(infection.bodyside)
exp(infection.bodyside$coefficient[2])
exp(infection.bodyside$coefficient[2]+c(-1,1)*1.96*0.3881)
infection.ANCcont<-glm(data=data,infection~ANCcont,family=binomial(link="logit"))
summary(infection.ANCcont)
exp(infection.ANCcont$coefficient[2])
exp(infection.ANCcont$coefficient[2]+c(-1,1)*1.96*0.05494)

infection.ANCcate<-glm(data=data,infection~ANCcate,family=binomial(link="logit"))
summary(infection.ANCcate)
exp(infection.ANCcate$coefficient[2])
exp(infection.ANCcate$coefficient[2]+c(-1,1)*1.96*0.36678)

infection.dysfun.ANCcate<-
glm(data=data,infection~factor(dysfunCVL)*factor(ANCcate),family=binomial(link="logit"))
summary(infection.dysfun.ANCcate)
```

7.2 SAS Code: Multiple Imputation Logistic Regression

```
proc import out= all_data datafile= "D:\trialsas1.xls"
dbms=excel2000 replace;
range="Sheet1$A1:O151";
getnames=yes;
run;

# with rounding adjustment
proc mi data=all_data seed=8633155 out=mimcmc
        round=1 1 1 1 1 1 1 1 1 0.1 1 1 noprint;
var TE infection agecont gender risk chemophase insertion dysfunCVL asprgtype ANCcont
typeCVL bodyside;
run;

data mimcmc;
        modify mimcmc;
        if TE>=0.5 then TE=1;
        else TE=0;
        if infection>=0.5 then infection=1;
        else infection=0;
        if agecont<0 then agecont=0;
        if agecont<=10 and agecont>1 then agecate=1;
        else agecate=0;
        if gender>=0.5 then gender=1;
        else gender=0;
        if risk>=0.5 then risk=1;
        else risk=0;
        if chemophase<0.5 then chemophase=0;
        else if chemophase>=0.5 and chemophase<1.5 then chemophase=1;
        else chemophase=2;
        if insertion>=0.5 then insertion=1;
        else insertion=0;
        if dysfunCVL>=0.5 then dysfunCVL=1;
        else dysfunCVL=0;
        if asprgtype>=0.5 then asprgtype=1;
        else asprgtype=0;
        if bodyside<0.5 then bodyside=0;
        else bodyside=1;
        if ANCcont<0 then ANCcont=0;
        if ANCcont<0.5 then ANCcate=1;
        else ANCcate=0;
run;

proc logistic data=mimcmc outest=outlg descending covout noprint;
class chemophase;
model TE =gender;
by _imputation_;
run;
```

```
proc print data=outlg;
        var _Imputation_ _Type_ _Name_ Intercept;
title 'Logistic Model Coefficients and Covariance Matrix from Imputed Data Sets';
run;

proc mianalyze data=outlg;
var Intercept gender;
run;

#without rounding adjustment
proc mi data=all_data seed=8633155 out=mimcmc noprint;
var TE infection agecont gender risk chemophase insertion dysfunCVL asprgtype ANCcont
typeCVL bodyside;
run;

data mimcmc (drop=typeCVL);
        modify mimcmc;
if TE<0.5 then TE=0;
        else TE=1;
if infection<0.5 then infection=0;
        else infection=1;
if chemophase<0.5 then chemophase=0;
        else if chemophase>=0.5 and chemophase<1.5 then chemophase=1;
        else chemophase=2;
if agecont<=10 and agecont>1 then agecate=1;
        else agecate=0;
if ANCcont<0.5 then ANCcate=1;
        else ANCcate=0;
run;

proc print data=mimcmc;
title 'Imputed Data Sets';
run;

ods listing close;
proc logistic data=mimcmc outest=outlg descending covout noprint;
class chemophase;
model TE =chemophase;
by _imputation_;
run;

proc mianalyze data=outlg;
var Intercept chemophase1 chemophase2;
run;

data mimcmc1;
        set mimcmc;
        if _imputation_ ne 1 then delete;
run;

data mimcmc2;
        set mimcmc;
        if _imputation_ ne 2 then delete;
```

```
run;

data mimcmc3;
        set mimcmc;
        if _imputation_ ne 3 then delete;
run;

data mimcmc4;
        set mimcmc;
        if _imputation_ ne 4 then delete;
run;

data mimcmc5;
        set mimcmc;
        if _imputation_ ne 5 then delete;
run;
```

7.3 R Code: Forest Plots

```
library (rmeta)
data<-read.table("ORs.txt", header=T)

text_te1<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                c("","OR","", data[1:4,2]),
                c("95%","Lower","", data[5:8,2]),
                c("95%","Upper","", data[9:12,2])
                 )
m<-c(NA,NA,NA,log(data[1:4,2]))
l<-c(NA,NA,NA,log(data[5:8,2]))
u<-c(NA,NA,NA,log(data[9:12,2]))
forestplot(text_te1,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Infection by TE Odds Ratio")

text_te2<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                c("","OR","", data[1:4,3]),
                c("95%","Lower","", data[5:8,3]),
                c("95%","Upper","", data[9:12,3])
                 )
m<-c(NA,NA,NA,log(data[1:4,3]))
l<-c(NA,NA,NA,log(data[5:8,3]))
u<-c(NA,NA,NA,log(data[9:12,3]))
forestplot(text_te2,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)),xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Age by TE Odds Ratio
(1-year increase)")


text_te3<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                c("","OR","", data[1:4,4]),
                c("95%","Lower","", data[5:8,4]),
                c("95%","Upper","", data[9:12,4])
                 )
m<-c(NA,NA,NA,log(data[1:4,4]))
l<-c(NA,NA,NA,log(data[5:8,4]))
u<-c(NA,NA,NA,log(data[9:12,4]))
forestplot(text_te3,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Age Category by TE Odds Ratio
(between 1 and 10 to otherwise)")


text_te4<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                c("","OR","", data[1:4,5]),
```

```
                    c("95%","Lower","", data[5:8,5]),
                    c("95%","Upper","", data[9:12,5])
                     )
m<-c(NA,NA,NA,log(data[1:4,5]))
l<-c(NA,NA,NA,log(data[5:8,5]))
u<-c(NA,NA,NA,log(data[9:12,5]))
forestplot(text_te4,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)),xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Gender by TE Odds Ratio
(female to male)")


text_te5<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                    c("","OR","", data[1:4,6]),
                    c("95%","Lower","", data[5:8,6]),
                    c("95%","Upper","", data[9:12,6])
                     )
m<-c(NA,NA,NA,log(data[1:4,6]))
l<-c(NA,NA,NA,log(data[5:8,6]))
u<-c(NA,NA,NA,log(data[9:12,6]))
forestplot(text_te5,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("ALL-Risk by TE Odds Ratio
(High Risk to Standard Risk)")


text_te6<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                    c("","OR","", data[1:4,7]),
                    c("95%","Lower","", data[5:8,7]),
                    c("95%","Upper","", data[9:12,7])
                     )
m<-c(NA,NA,NA,log(data[1:4,7]))
l<-c(NA,NA,NA,NA,log(data[5:8,7]))
u<-c(NA,NA,NA,log(data[9:12,7]))
forestplot(text_te6,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("At-Insertion Chemophase by TE Odds Ratio
(before chemotherapy to post chemotherapy)")


text_te7<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                    c("","OR","", data[1:4,8]),
                    c("95%","Lower","", data[5:8,8]),
                    c("95%","Upper","", data[9:12,8])
                     )
m<-c(NA,NA,NA,log(data[1:4,8]))
l<-c(NA,NA,NA,log(data[5:8,8]))
u<-c(NA,NA,NA,log(data[9:12,8]))
forestplot(text_te7,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
```

```
title ("At-Insertion Chemophase by TE Odds Ratio
(induction to post chemotherapy)")


text_te8<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                        c("","OR","", data[1:4,9]),
                        c("95%","Lower","", data[5:8,9]),
                        c("95%","Upper","", data[9:12,9])
                         )
m<-c(NA,NA,NA,log(data[1:4,9]))
l<-c(NA,NA,NA,log(data[5:8,9]))
u<-c(NA,NA,NA,log(data[9:12,9]))
forestplot(text_te8,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("CVL Insertion by TE Odds Ratio
(percutaneous to open)")


text_te9<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic MI
w/o Rounding"),
                        c("","OR","", data[1:4,10]),
                        c("95%","Lower","", data[5:8,10]),
                        c("95%","Upper","", data[9:12,10])
                         )
m<-c(NA,NA,NA,log(data[1:4,10]))
l<-c(NA,NA,NA,log(data[5:8,10]))
u<-c(NA,NA,NA,log(data[9:12,10]))
forestplot(text_te9,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("CVL Functionality by TE Odds Ratio
(dysfunction to function)")


text_te10<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic MI&Rounding","Logistic
MI w/o Rounding"),
                        c("","OR","", data[1:4,11]),
                        c("95%","Lower","", data[5:8,11]),
                        c("95%","Upper","", data[9:12,11])
                         )
m<-c(NA,NA,NA,log(data[1:4,11]))
l<-c(NA,NA,NA,log(data[5:8,11]))
u<-c(NA,NA,NA,log(data[9:12,11]))
forestplot(text_te10,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Site of Insertion by TE Odds Ratio
(left to right)")


data<-read.table("ORs.txt", header=T)

text_infection1<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
```

```
                        c("","OR","", data[1:4,12]),
                        c("95%","Lower","", data[5:8,12]),
                        c("95%","Upper","", data[9:12,12])
                         )
m<-c(NA,NA,NA,NA,log(data[1:4,12]))
l<-c(NA,NA,NA,NA,log(data[5:8,12]))
u<-c(NA,NA,NA,NA,log(data[9:12,12]))
forestplot(text_infection1,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), ,xlog=TRUE, col=meta.colors(box="royalblue",line="darkblue"))
title ("TE by Infection Odds Ratio")

text_infection2<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                        c("","OR","", data[1:4,13]),
                        c("95%","Lower","", data[5:8,13]),
                        c("95%","Upper","", data[9:12,13])
                         )
m<-c(NA,NA,NA,log(data[1:4,13]))
l<-c(NA,NA,NA,log(data[5:8,13]))
u<-c(NA,NA,NA,log(data[9:12,13]))
forestplot(text_infection2,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Age by Infection Odds Ratio
(1-year increase)")

text_infection3<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                        c("","OR","", data[1:4,14]),
                        c("95%","Lower","", data[5:8,14]),
                        c("95%","Upper","", data[9:12,14])
                         )
m<-c(NA,NA,NA,log(data[1:4,14]))
l<-c(NA,NA,NA,log(data[5:8,14]))
u<-c(NA,NA,NA,log(data[9:12,14]))
forestplot(text_infection3,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Age Category by Infection Odds Ratio
(between 1 and 10 to otherwise)")

text_infection4<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                        c("","OR","", data[1:4,15]),
                        c("95%","Lower","", data[5:8,15]),
                        c("95%","Upper","", data[9:12,15])
                         )
m<-c(NA,NA,NA,log(data[1:4,15]))
l<-c(NA,NA,NA,log(data[5:8,15]))
u<-c(NA,NA,NA,log(data[9:12,15]))
forestplot(text_infection4,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("Gender by Infection Odds Ratio
(female to male)")
```

```
text_infection5<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                       c("","OR","", data[1:4,16]),
                       c("95%","Lower","", data[5:8,16]),
                       c("95%","Upper","", data[9:12,16])
                        )
m<-c(NA,NA,NA,log(data[1:4,16]))
l<-c(NA,NA,NA,log(data[5:8,16]))
u<-c(NA,NA,NA,log(data[9:12,16]))
forestplot(text_infection5,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("ALL-Risk by Infection Odds Ratio
(High Risk to Standard Risk)")

text_infection6<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                       c("","OR","", data[1:4,17]),
                       c("95%","Lower","", data[5:8,17]),
                       c("95%","Upper","", data[9:12,17])
                        )
m<-c(NA,NA,NA,log(data[1:4,17]))
l<-c(NA,NA,NA,log(data[5:8,17]))
u<-c(NA,NA,NA,log(data[9:12,17]))
forestplot(text_infection6,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("At-Insertion Chemophase by Infection Odds Ratio
(before chemotherapy to post chemotherapy)")

text_infection7<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                       c("","OR","", data[1:4,18]),
                       c("95%","Lower","", data[5:8,18]),
                       c("95%","Upper","", data[9:12,18])
                        )
m<-c(NA,NA,NA,log(data[1:4,18]))
l<-c(NA,NA,NA,log(data[5:8,18]))
u<-c(NA,NA,NA,log(data[9:12,18]))
forestplot(text_infection7,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("At-Insertion Chemophase by Infectioin Odds Ratio
(induction to others)")

text_infection8<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                       c("","OR","", data[1:4,19]),
                       c("95%","Lower","", data[5:8,19]),
                       c("95%","Upper","", data[9:12,19])
                        )
m<-c(NA,NA,NA,log(data[1:4,19]))
l<-c(NA,NA,NA,log(data[5:8,19]))
u<-c(NA,NA,NA,log(data[9:12,19]))
forestplot(text_infection8,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
```

```
title ("CVL Insertion by Insertion Odds Ratio
(percutaneous to open)")

text_infection9<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                       c("","OR","", data[1:4,20]),
                       c("95%","Lower","", data[5:8,20]),
                       c("95%","Upper","", data[9:12,20])
                        )
m<-c(NA,NA,NA, log(data[1:4,20]))
l<-c(NA,NA,NA,log(data[5:8,20]))
u<-c(NA,NA,NA,log(data[9:12,20]))
forestplot(text_infection9,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("CVL Functionality by Infection Odds Ratio
(dysfunction to function)")

text_infection10<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                       c("","OR","", data[1:4,21]),
                       c("95%","Lower","", data[5:8,21]),
                       c("95%","Upper","", data[9:12,21])
                        )
m<-c(NA,NA,NA,log(data[1:4,21]))
l<-c(NA,NA,NA,log(data[5:8,21]))
u<-c(NA,NA,NA,log(data[9:12,21]))
forestplot(text_infection10,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("ANC by Infection Odds Ratio
(per 10^9 / L increase)")

text_infection11<-cbind(c("","Method","","Logistic w/o MI","Bayesian","Logistic
MI&Rounding","Logistic MI w/o Rounding"),
                       c("","OR","", data[1:4,22]),
                       c("95%","Lower","", data[5:8,22]),
                       c("95%","Upper","", data[9:12,22])
                        )
m<-c(NA,NA,NA,log(data[1:4,22]))
l<-c(NA,NA,NA,log(data[5:8,22]))
u<-c(NA,NA,NA,log(data[9:12,22]))
forestplot(text_infection11,m,l,u,zero=0,is.summary=c(TRUE,TRUE,rep(FALSE,5)),
clip=c(log(0.01),log(10)), xlog=TRUE,col=meta.colors(box="royalblue",line="darkblue"))
title ("ANC Category by Infection Odds Ratio
(less than 0.5 x 10^9 / L to others)")
```

7.4 SAS Code: Fisher's Exact Test

```
data pred;
input freq predictor $ outcome $;
datalines;
15 Y Y
67 N Y
 7 Y N
54 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
15 Y Y
7 N Y
99 Y N
25 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
11 Y Y
11 N Y
51 Y N
73 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
10 Y Y
```

```
12 N Y
50 Y N
73 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
14 Y Y
 1 N Y
59 Y N
 6 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
 9 Y Y
 1 N Y
59 Y N
 6 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
 7 Y Y
14 N Y
59 Y N
59 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
```

```
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
14 Y Y
 8 N Y
84 Y N
35 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
 7 Y Y
14 N Y
27 Y N
94 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;


data pred;
input freq predictor $ outcome $;
datalines;
 3 Y Y
19 N Y
34 Y N
87 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
15 Y Y
```

```
 0 N Y
80 Y N
23 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
 19 Y Y
  3 N Y
121 Y N
  2 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
61 Y Y
22 N Y
52 Y N
10 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
36 Y Y
47 N Y
26 Y N
36 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
```

```
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
46 Y Y
36 N Y
37 Y N
25 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
41 Y Y
 2 N Y
30 Y N
 5 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
40 Y Y
 2 N Y
27 Y N
 5 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
40 Y Y
41 N Y
```

```
27 Y N
30 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
55 Y Y
25 N Y
41 Y N
19 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
26 Y Y
54 N Y
 9 Y N
51 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;

data pred;
input freq predictor $ outcome $;
datalines;
77 Y Y
 6 N Y
62 Y N
 0 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
```

```
run;

data pred;
input freq predictor $ outcome $;
datalines;
38 Y Y
41 N Y
17 Y N
40 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;


data pred;
input freq predictor $ outcome $;
datalines;
19 Y Y
62 N Y
17 Y N
43 N N
;
proc freq data=pred; weight freq;
tables predictor*outcome;
exact fisher or/alpha=0.05; run;
proc logistic data=pred; freq freq; class predictor outcome;
model outcome=predictor/clodds=pl;
run;
```

7.5 Winbugs Code: Bayesian Analysis

```
#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.infection*infection[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.infection~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.infection=1)
list(n = 143)
TE[]    infection[]
1       1
1       1
...
END

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.agecont*agecont[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.agecont ~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.agecont=1)
list(n = 146)
TE[]    agecont[]
1       13
1       4
0       10
...
END

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.agecate*agecate[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.agecate ~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.agecate=1)
list(n = 146)
```

```
TE[]    agecate[]
1       0
1       1
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.gender*gender[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.gender ~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.gender=1)
list(n = 146)
TE[]    gender[]
1       1
1       0
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.risk*risk[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.risk~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.risk=1)
list(n = 145)
TE[]    risk[]
1       1
1       0
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.chemophase*chemophase[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.chemophase~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.chemophase=1)
```

```
list(n = 146)
TE[]    chemophase[]
1       1
1       1
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.insertion*insertion[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.insertion~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.insertion=1)
list(n = 141)
TE[]    insertion[]
1       1
1       0
...
End
#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.dysfunCVL*dysfunCVL[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.dysfunCVL~ dnorm(0.0,1.0E-4);
}

#Initial Values
list(alpha=0, b.dysfunCVL=1)
list(n = 142)
TE[]    dysfunCVL[]
1       0
1       1
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.bodyside*bodyside[i];
TE[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.bodyside~ dnorm(0.0,1.0E-4);
}
```

```
#Initial Values
list(alpha=0, b.bodyside=1)

list(n = 143)
TE[]    bodyside[]
1       1
1       0
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.te*TE[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.te~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.te=1)
list(n = 143)
infection[]     TE[]
1       1
1       1
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.agecont*agecont[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.agecont~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.agecont=1)
list(n = 145)
infection[]     agecont[]
1       13
1       4
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.agecate*agecate[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
```

```
b.agecate~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.agecate=1)
list(n = 145)
infection[]        agecate[]
1        0
1        1
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.gender*gender[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.gender~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.gender=1)
list(n = 145)
infection[]        gender[]
1        1
1        0
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.risk*risk[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.risk~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.risk=1)
list(n = 144)
infection[]        risk[]
1        1
1        0
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.chemophase*chemophase[i];
infection[i] ~ dbern(p[i]);
}
```

```
alpha ~ dnorm(0.0,1.0E-4);
b.chemophase~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.chemophase=1)
list(n = 145)
infection[]        chemophase[]
1          1
1          1
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.insertion*insertion[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.insertion~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.insertion=1)
list(n = 140)
infection[]        insertion[]
1          1
1          0
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.dysfunCVL*dysfunCVL[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.dysfunCVL~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.dysfunCVL=1)
list(n = 140)
infection[]        dysfunCVL[]
1          0
1          1
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.ANCcont*ANCcont[i];
infection[i] ~ dbern(p[i]);
```

```
}
alpha ~ dnorm(0.0,1.0E-4);
b.ANCcont~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.ANCcont=1)
list(n = 136)
infection[]        ANCcont[]
1        5.4
1        5
...
End

#Model (Logistic Regression)
model {
for (i in 1:n) {
logit(p[i]) <- alpha + b.ANCcate*ANCcate[i];
infection[i] ~ dbern(p[i]);
}
alpha ~ dnorm(0.0,1.0E-4);
b.ANCcate~ dnorm(0.0,1.0E-4);
}
#Initial Values
list(alpha=0, b.ANCcate=1)
list(n = 136)
infection[]        ANCcate[]
1        0
1        0
...
End
```

7.6 S-Plus Code: CART Modeling

```
te.trialsas1<-tree(formula = TE ~ infection + agecate + gender + risk + chemophase + insertion +
dysfunCVL + bodyside,data = trialsas1, na.action = na.exclude, mincut = 5, minsize = 10, mindev
= 0.01)
summary(te.trialsas1)
summary(prune.tree(te.trialsas1,best=2))
plot(prune.tree(te.trialsas1, best=2))
text(prune.tree(te.trialsas1, best=2))
title("Classification Tree for TE with the Original Dataset
3 Nodes, 15.79% Misclass,  0.8278 Res. Mean Deviance")

infection.trialsas1<-tree(formula = infection ~ TE + agecate + gender + risk + chemophase +
insertion + dysfunCVL + ANCcate,data = trialsas1, na.action = na.exclude, mincut = 5, minsize =
10, mindev = 0.01)
summary(infection.trialsas1)
summary(prune.tree(infection.trialsas1,best=5))
plot(prune.tree(infection.trialsas1, best=5))
text(prune.tree(infection.trialsas1, best=5))
title("Classification Tree for Infection with the Original Dataset
7 Nodes,  29.37% Misclass,  1.257 Res. Mean Deviance")

te.mimcmc1<-tree(formula = TE ~ infection + agecate + gender + risk + chemophase + insertion
+ dysfunCVL + bodyside,        data = mimcmc1, na.action = na.exclude, mincut = 5, minsize =
10, mindev = 0.01)
summary(te.mimcmc1)
plot(te.mimcmc1)
text(te.mimcmc1)
title("Classification Tree for TE with the 1st Imputed Dataset
18 Nodes,  15.33% Misclass,  0.7704 Res. Mean Deviance")
summary(prune.tree(te.mimcmc1,best=2))
plot(prune.tree(te.mimcmc1,best=2))
text(prune.tree(te.mimcmc1,best=2))
title("Classification Tree for TE with the 1st Imputed Dataset
4 Nodes,  15.33% Misclass,  0.8139 Res. Mean Deviance")

infection.mimcmc1<-tree(formula = infection ~ TE + agecate + gender + risk + chemophase +
insertion + dysfunCVL + ANCcate,data = mimcmc1, na.action = na.exclude, mincut = 5, minsize
= 10, mindev = 0.01)
summary(infection.mimcmc1)
summary(prune.tree(infection.mimcmc1,best=3))
plot(prune.tree(infection.mimcmc1,best=3))
text(prune.tree(infection.mimcmc1,best=3))
title("Classification Tree for Infection with the 1st Imputed Dataset
8 Nodes,  32% Misclass,  1.266 Res. Mean Deviance")

te.mimcmc2<-tree(formula = TE ~ infection + agecate + gender + risk + chemophase + insertion
+ dysfunCVL + bodyside,data = mimcmc2, na.action = na.exclude, mincut = 5, minsize = 10,
mindev = 0.01)
summary(te.mimcmc2)
summary(prune.tree(te.mimcmc2,best=2))
plot(prune.tree(te.mimcmc2, best=2))
```

```
text(prune.tree(te.mimcmc2, best=2))
title("Classification Tree for TE with the 2nd Imputed Dataset
4 Nodes,  14.67% Misclass,  0.7633 Res. Mean Deviance")

infection.mimcmc2<-tree(formula = infection ~ TE + agecate + gender + risk + chemophase +
insertion + dysfunCVL +ANCcate, data = mimcmc2, na.action = na.exclude, mincut = 5, minsize
= 10, mindev = 0.01)
summary(infection.mimcmc2)
summary(prune.tree(infection.mimcmc2,best=5))
plot(prune.tree(infection.mimcmc2, best=5))
text(prune.tree(infection.mimcmc2, best=5))
title("Classification Tree for Infection with the 2nd Imputed Dataset
5 Nodes,  32.67% Misclass,  1.285 Res. Mean Deviance")

te.mimcmc3<-tree(formula = TE ~ infection + agecate + gender + risk + chemophase + insertion
+ dysfunCVL + bodyside,data = mimcmc3, na.action = na.exclude, mincut = 5, minsize = 10,
mindev = 0.01)
summary(te.mimcmc3)
summary(prune.tree(te.mimcmc3,best=5))
plot(prune.tree(te.mimcmc3, best=5))
text(prune.tree(te.mimcmc3, best=5))
title("Classification Tree for TE with the 3rd Imputed Dataset
5 Nodes, 15.33% Misclass,  0.7831 Res. Mean Deviance")

infection.mimcmc3<-tree(formula = infection ~ TE + agecate + gender + risk + chemophase +
insertion + dysfunCVL + ANCcate,data = mimcmc3, na.action = na.exclude, mincut = 5, minsize
= 10, mindev = 0.01)
summary(infection.mimcmc3)
summary(prune.tree(infection.mimcmc3, best=8))
plot(prune.tree(infection.mimcmc3, best=8))
text(prune.tree(infection.mimcmc3, best=8))
title("Classification Tree for Infection with the 3rd Imputed Dataset
9 Nodes,  32% Misclass,  1.278 Res. Mean Deviance")

te.mimcmc4<-tree(formula = TE ~ infection + agecate + gender + risk + chemophase + insertion
+ dysfunCVL + bodyside,data = mimcmc4, na.action = na.exclude, mincut = 5, minsize = 10,
mindev = 0.01)
summary(te.mimcmc4)
summary(prune.tree(te.mimcmc4,best=2))
plot(prune.tree(te.mimcmc4, best=2))
text(prune.tree(te.mimcmc4, best=2))
title("Classification Tree for TE with the 4th Imputed Dataset
4 Nodes, 15.33% Misclass,  0.7836 Res. Mean Deviance")

infection.mimcmc4<-tree(formula = infection ~ TE + agecate + gender + risk + chemophase +
insertion + dysfunCVL + ANCcate,data = mimcmc4, na.action = na.exclude, mincut = 5, minsize
= 10, mindev = 0.01)
summary(infection.mimcmc4)
summary(prune.tree(infection.mimcmc4,best=5))
plot(prune.tree(infection.mimcmc4, best=5))
text(prune.tree(infection.mimcmc4, best=5))
title("Classification Tree for Infection with the 4th Imputed Dataset
8 Nodes,  32.67% Misclass,  1.282 Res. Mean Deviance")
```

```
te.mimcmc5<-tree(formula = TE ~ infection + agecate + gender + risk + chemophase + insertion
+ dysfunCVL + bodyside,data = mimcmc5, na.action = na.exclude, mincut = 5, minsize = 10,
mindev = 0.01)
summary(te.mimcmc5)
summary(prune.tree(te.mimcmc5,best=2))
plot(prune.tree(te.mimcmc5, best=2))
text(prune.tree(te.mimcmc5, best=2))
title("Classification Tree for TE with the 5th Imputed Dataset
3 Nodes, 15.33% Misclass,   0.8175 Res. Mean Deviance")

infection.mimcmc5<-tree(formula = infection ~ TE + agecate + gender + risk + chemophase +
insertion + dysfunCVL + ANCcate,data = mimcmc5, na.action = na.exclude, mincut = 5, minsize
= 10, mindev = 0.01)
summary(infection.mimcmc5)
summary(prune.tree(infection.mimcmc5,best=4))
plot(prune.tree(infection.mimcmc5, best=4))
text(prune.tree(infection.mimcmc5, best=4))
title("Classification Tree for Infection with the 5th Imputed Dataset
6 Nodes, 32.67% Misclass,   1.3 Res. Mean Deviance")
```

7.7 SPSS 14 Code: CART Modeling

```
* TE
* CHAID (Pearson)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES
CHISQUARE=PEARSON CONVERGE=0.001
    MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* TE
* CHAID (Likelihood)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES CHISQUARE=LR
CONVERGE=0.001
    MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* TE
* CRT (Gini)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
```

```
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=GINI MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.


* TE
* CRT (Entropy)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=TWOING MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.


* TE
* QUEST
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=QUEST MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /QUEST ALPHASPLIT=0.2
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.
```

```
* TE
* CHAID(Pearson)
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES
CHISQUARE=PEARSON CONVERGE=0.001 MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* TE
* CHAID(Likelihood)
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES CHISQUARE=LR
CONVERGE=0.001 MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* TE
* CRT(Gini)
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
```

```
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=GINI MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.


* TE
* CRT(Twoing)
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=TWOING MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.


* TE
* QUEST
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE TE [n] BY agecate [n] gender [n] risk [n] infection [n] chemophase [n] bodyside [n] insertion
[n] dysfunCVL [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=QUEST MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /QUEST ALPHASPLIT=0.2
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING



* INFECTION
* CHAID (Pearson)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
```

```
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES
CHISQUARE=PEARSON CONVERGE=0.001
    MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* INFECTION
* CHAID (Likelihood)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES CHISQUARE=LR
CONVERGE=0.001
    MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* INFECTION
* CRT (Gini)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=GINI MINIMPROVEMENT=0.0001
  /COSTS EQUAL
```

```
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.

* INFECTION
* CRT (Entropy)
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=TWOING MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.


* INFECTION
* QUEST
* SPLIT SAMPLING (Training 90%, Test 10%)

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=QUEST MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=SPLITSAMPLE(90) OUTPUT=BOTHSAMPLES
  /QUEST ALPHASPLIT=0.2
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.


* INFECTION
* CHAID (Pearson)
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
```

```
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES
CHISQUARE=PEARSON CONVERGE=0.001
    MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* INFECTION
* CHAID (Likelihood)
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CHAID
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.2 ALPHAMERGE=0.2 SPLITMERGED=YES CHISQUARE=LR
CONVERGE=0.001 MAXITERATIONS=100 ADJUST=BONFERRONI
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.


* INFECTION
* CRT (Gini)
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=GINI MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.


* INFECTION
* CRT (Gini)
```

```
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=TWOING MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.



* INFECTION
* QUEST
* 10-FOLD CROSS VALIDATION

* Classification Tree.
TREE infection [n] BY agecate [n] gender [n] risk [n] chemophase [n] bodyside [n] insertion [n]
ANCcate [n] dysfunCVL [n] TE [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=QUEST MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=5 MINPARENTSIZE=10 MINCHILDSIZE=5
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /QUEST ALPHASPLIT=0.2
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISS
```