Original Investigations

JAMIA

SCAMC Student Paper Competition Finalist ■

An Evaluation of Factors Influencing Bayesian Learning Systems

ERIC L. EISENSTEIN, MBA, FARROKH ALEMI, PhD

Abstract **Objectives:** To examine the influences of situational and model factors on the accuracy of Bayesian learning systems.

Design: This study examines the impacts of variations in two situational factors, training sample size and number of attributes, and in two model factors, choice of Bayesian model and criteria for excluding model attributes, on the overall accuracy of Bayesian learning systems.

Measurements: The test data were derived from myocardial infarction patients who were admitted to eight hospitals in New Orleans during 1985. The test sample consisted of 339 cases; the training samples included 100, 400, and 800 cases. APACHE II variables were used for the model attributes and patient discharge status as the outcome predicted. Attribute sets were selected in sizes of 4, 8, and 12. The authors varied the Bayesian models (proper and simple) and the attribute exclusion criteria (optimism and pessimism).

Results: The simple Bayes model, which assumes conditional independence, consistently equalled or outperformed the proper (maximally dependent) Bayes model, which assumes conditional dependence, across all training sample and attribute set sizes. Not excluding model attributes was found to be preferable to using sample theory as an attribute exclusion criterion in both the simple and the proper models.

Conclusion: In the domain tested, the simple Bayes model with optimistic exclusion is more robust than previously assumed and increasing the number of attributes in a model had a greater relative impact on model accuracy than did increasing the number of training sample cases. Assessment of applicability of these findings to other domains will require further study. In addition, other models that are between these two extremes must be investigated. These include models that approximate proper Bayes' conditional dependence computations while requiring fewer training sample cases, attribute exclusion criteria between optimism and pessimism that improve accuracy, and ordering techniques for introducing attributes into Bayes models that optimize the information value associated with the attributes in test-sample cases.

■ J Am Med Informatics Assoc. 1994;1:272-284.

Affiliation of the authors: Cleveland State University, Cleveland, OH.

Correspondence and reprints: Eric L. Eisenstein, 6812 Old Royalton Road, #3, Brecksville, OH 44141.

Presented in part at the Symposium on Computer Applications in Medical Care, 1993.

Received for publication: 11/4/93; accepted for publication: 11/11/93.

"Conditional independence means that for a specific population, the presence of one characteristic does not change the value of any other."¹ As an example, if the population of interest is patients admitted to a hospital and age and chronic health problem are the characteristics, conditional independence would mean that, for this population, the probability of a patient's

being over 65 years of age, given that he or she is admitted to a hospital, is equal to the probability that he or she is over 65 years of age, given that he or

she is admitted to a hospital and has a chronic health problem. The formula for this relationship is shown below.

$$p(\text{age} > 65 | \text{admitted}) = p(\text{age} > 65 | \text{admitted, chronic health problem})$$

Thus, conditional independence in this instance means that the probability that the patient is over 65 years of age is not related to whether or not the patient also has a chronic health problem. Clearly, many people would assert that age is related to chronic health problems and that it is often the case that the presence of certain patient characteristics may change the likelihoods of the values of other patient characteristics.

During the past two decades, there has been considerable discussion regarding the importance of the conditional-independence assumption in Bayesian analysis. One group of researchers has stated that the problems encountered in managing conditional dependence have been an impediment to the acceptance of Bayesian analysis.²⁻⁵ In contrast, other researchers have asserted that the simple Bayes model is relatively robust and that the management of conditional dependence is not an important problem.⁶⁻⁹ This study reviews those factors that have been assumed to influence the importance of the conditional-independence assumption in Bayesian analysis. It also attempts to identify situations in which Bayesian models that assume conditional independence and Bayesian models that assume conditional dependence in their data are, respectively, more appropriate.

Bayesian Terminology

Bayesian analysis is often concerned with predicting whether one of two mutually exclusive and collectively exhaustive events will occur. The probability of such an event's occurring, without considering the presence of additional data, is called the prior prob-

ability. The probability of such an event's occurring, considering the presence of additional data, is called the posterior probability.

The odds of an event are defined as the probability that the event will occur divided by the probability that the event will not occur. Thus, the prior odds of an event's occurring are the odds without considering the effect of an additional set of data and the posterior odds of an event's occurring are the odds after considering the effect of an additional set of data. Bayes first suggested a formula for computing posterior probabilities.¹⁰ The posterior odds version of this formula is given below. In this notation, the A_i s are the individual attribute values and H is the hypothesis of interest. H_1 denotes the hypothesis is true and H_0 denotes the hypothesis is false.

$$\text{Posterior odds of } H = \frac{p(A_1, A_2, \dots, A_n | H_1)}{p(A_1, A_2, \dots, A_n | H_0)} \times \frac{P(H_1)}{P(H_0)}$$

This formula states that the posterior odds of event H are equal to the product of two terms. The first term is the combined likelihood ratios for all attributes in the set of additional data that are used to predict event H . The second term is the prior odds of event H before the effect of the additional set of data is considered.

The likelihood ratio for each attribute is the probability that the attribute has a specific value given that event H occurs divided by the probability of the attribute's having the same value given that event H does not occur. If the data are assumed to be conditionally independent, the likelihood ratios in the posterior odds formula can be expanded as shown below.

$$\frac{p(A_1, A_2, \dots, A_n | H_1)}{p(A_1, A_2, \dots, A_n | H_0)} = \frac{p(A_1 | H_1)}{p(A_1 | H_0)} \times \frac{p(A_2 | H_1)}{p(A_2 | H_0)} \times \dots \times \frac{p(A_n | H_1)}{p(A_n | H_0)}$$

If the conditional-independence assumption is not valid and the attributes are interdependent, the likelihood ratio formula must be modified to account for

joint likelihood ratios in the set of attributes. This revised formula is shown below.

$$\frac{p(A_1, A_2, \dots, A_n|H_1)}{p(A_1, A_2, \dots, A_n|H_0)} = \frac{p(A_1|H_1)}{p(A_1|H_0)} \times \frac{p(A_2|H_1, A_1)}{p(A_2|H_0, A_1)} \times \dots \times \frac{p(A_n|H_1, A_1, A_2, \dots, A_{n-1})}{p(A_n|H_0, A_1, A_2, \dots, A_{n-1})}$$

Each term in this formula is conditioned on the values for all of the previous terms. Only the first term is written in its conditionally independent form. All other terms are in their conditionally nonindependent forms.

These two formulas define the primary Bayesian models that have been used in previous studies. Although they have many names, below we call the formula that assumes conditional independence the "simple" Bayes model and the formula that assumes conditional dependence the "proper" Bayes model.¹¹

Situational Factors

Situational factors are characteristics of training samples that may influence the relative efficacy of the simple and proper Bayes models. The presence or absence of these factors, and their interactions, are often assumed to determine which of the Bayes models will have the greater accuracy. Three situational factors are prominent in the literature. These are: 1) conditional independence/dependence of model attributes, 2) training sample size, and 3) the number of model attributes. These situational factors and the research design questions they raise are discussed below.

Conditional Dependence

The existence of conditional dependence in training-data sets has a direct bearing upon the choice between the simple and proper Bayes models. If data are conditionally independent, there are no joint likelihood ratios and the simple Bayes model can be used for computations. With this model, the likelihood ratios for each attribute are computed from the entire training sample. Thus, if there were 800 cases in the training data set, all 800 cases would be used in computing the likelihood ratios for each attribute. In contrast, when data are conditionally nonindependent, joint likelihood ratios exist between the attributes and the proper Bayes model should be used

to calculate the conditional likelihood ratios. These conditional likelihood ratios are computed by successively reducing the training sample, after each attribute's likelihood ratios are computed, to account for conditional dependencies with attributes that were previously included in the model.

When the proper Bayes model is used to compute likelihood ratios, the conditional likelihood ratio for A_1 , the first attribute in the model, is computed from the entire training sample and the conditional likelihood ratio for A_2 , the second attribute in the model, is computed from only that portion of the training sample that shares the same value for A_1 . For example, if the response to A_1 is patient age greater than 65 years and the response to A_2 is patient has a chronic health problem, then the likelihood ratio for A_1 would be computed from all cases in the training sample and that for A_2 would be computed from only those cases in the training sample where the patients' ages were more than 65 years. If the training sample contains 800 cases and 50% of the patient are more than 65 years of age, the ratio for A_1 will be computed from 800 cases and that for A_2 will be computed from 400 cases.

Theoretically, the simple Bayes model should be used only when there is conditional independence in the training sample and the proper Bayes model should be used whenever conditional dependence exists. However, this rule has practical exceptions. These exceptions relate to limitations imposed by the size of the training sample or by the number of attributes that prevent full exercise of the proper Bayes model. With a small training sample in relation to the number of attributes, the proper Bayes model might run out of training sample cases prior to including all of the attributes. When this occurs, inaccuracy is introduced into the proper Bayes model's computations.

Several studies have evaluated the importance of conditional dependence upon the accuracy of the simple Bayes model; however, there were some lim-

itations to these studies. In the studies of Spicer⁷ and Weintraub et al.,⁹ small attribute sets were used. These have been associated with small differences between the accuracies of the simple and the proper Bayes models.¹²⁻¹⁴ Chard's study used seven attributes and eight outcomes. He found that within reasonable limits, dependence does not greatly affect outcome measurements. Chard also found that some of his most dramatic results were artifacts of his experiment and not attributable to conditional dependence. Other researchers have reported that the use of the simple Bayes model when there is conditional dependence in the attributes may yield results that are equal to or better than those of the proper Bayes model.⁸

Training Sample Size

Training sample size is generally considered to be a primary determinant of accuracy in both the simple and the proper Bayes models. However, there has been little research that substantiates this assumption. Chard generated conditionally independent training samples containing seven attributes and eight outcomes. The prior probabilities for this sample ranged from 2% to 30%. Chard concluded that a minimum of 200 cases is required for the simple Bayes model and that there is a maximum of approximately 500 cases above which increasing the training sample size will produce no further improvement in the model's accuracy.¹⁵

No study similar to Chard's but using conditionally dependent training samples has been published. However, Gammernan and Thatcher, based on a study with nine outcomes and 33 attributes, reported that a database containing 4,387 patients was not large enough to permit all relevant combinations of symptoms to be identified in adequate detail such that the proper Bayes model would outperform the simple Bayes model.¹¹ Thus, while 500 cases is the maximum number required to predict eight outcomes from seven attributes under conditional independence, 4,387 cases are insufficient to allow the proper Bayes model to outperform the simple Bayes model in predicting nine outcomes from 33 attributes under conditional dependence.

Fryback concluded that computing first-order conditional likelihood ratios (likelihood ratios conditioned only on the first preceding attribute), regardless of whether the conditional-independence assumption is satisfied, is valuable only when there is a sufficiently large training sample to reliably estimate the conditional dependencies in the proper Bayes model.¹³ Otherwise, the simple Bayes model should be used. While Fryback commented on first-

Table 1 ■

Situational Factors and Model Selection*

Factors and Models	Scenario			
	1	2	3	4
Situational factors				
Conditional dependence	Y	Y	Y	Y
Large training sample	Y	Y	N	N
Large number of attributes	Y	N	Y	N
Model choices				
Proper Bayes	X			
Simple Bayes			X	X
Neither model preferred		X		

*Y = yes; N = no.

order conditioning only, his observations could easily be expanded to all orders of conditioning, from 1 to $n - 1$. Thus, in choosing between these two Bayesian models, the proper Bayes model would be presumed to be more accurate than the simple Bayes model when there was conditional dependence in the attributes and when the training sample is large enough to reliably support all of its computations. When these conditions were not met, the simple Bayes model would be presumed to be at least as accurate as the proper Bayes model.

Number of Attributes

Several researchers have reported small differences in accuracy between the different Bayes models with small attribute sets and larger differences in accuracy with larger attribute sets.¹³⁻¹⁵ This led Fryback to question the value of using the proper Bayes model when there was a small attribute set. He hypothesized that with a small attribute set, the contribution of each attribute is stronger than the degradation in performance that results from errors caused by ignoring the dependencies between attributes. Thus, the proper Bayes model is favored when there is conditional dependence, a large training sample, and a large attribute set. In all other situations, where there is conditional dependence in the attributes, a large training sample, and a small number of attributes, there should be little difference in accuracy between the simple and proper Bayes models.

Situational-factors Summary

Table 1 summarizes the previous research regarding the influences of situational factors upon the choice between the simple and proper Bayes models. The proper Bayes model is assumed to be more accurate than the simple Bayes model in only one scenario.

This is scenario 1, where there is conditional dependence, a large training sample, and a large number of attributes. There is no assumed difference in accuracy between these two models when there is conditional dependence, a large training sample, and a small number of attributes. Scenarios 3 and 4 do not contain a large training sample, and the simple Bayes model should be more accurate than the proper Bayes model in these situations. Other possible scenarios involving conditional independence are not depicted in Table 1 as the simple Bayes model is always chosen in these situations.

Model Factors

Model factors influence the relative accuracies of the simple and proper Bayes models through their management of a model's data. The presence or absence of model factors can either enhance or diminish the relative performances of Bayesian models. Three factors are often mentioned in the literature. These are: 1) whether attributes are excluded from the final model, 2) what criteria are used to exclude attributes from the model, and 3) how attributes are ordered for introduction into the model. These model factors and the research design questions they raise are discussed below.

Attribute Exclusion

Ohmann et al.¹⁴ reported that none of the conditionally nonindependent models they tested achieved maximum diagnostic accuracy when all attributes were included. This led them to conclude that better results could be obtained if adequate strategies were used for the selection of attributes.

These authors also observed a relationship between the number of attributes and the training sample size in models that partition the training sample during computations. When these models do not provide for the exclusion of attributes, their accuracy tends to peak. That is, the model's accuracy will often improve up to a point and then begin to deteriorate as new attributes are added. This phenomenon is especially pronounced when the ratio of the training sample size to the number of attributes is small. Peaking also was a problem for other statistical models that partitioned their data sets and automatically included all attributes. By implication, peaking would also be a problem for the proper Bayes model, if provisions were not made for attribute exclusion.

Ohmann et al. also reported that monotonicity could be demonstrated in a simple Bayes model with many attributes, and they concluded that it may be pref-

erable to use all attributes in simple Bayes models. The reason for this is that the simple Bayes model, unlike the proper Bayes and many statistical models, does not partition the training sample during computations.

Exclusion Criteria

Both chi-square and sample-theory tests have been used in machine learning systems as measurements for attribute exclusion.^{16,17} However, as Fisher¹⁸ first noted, the real issue may not be the measurement that is used, but rather the confidence level that is chosen for excluding attributes.

Fisher observed that a pessimistic bias exists in machine learning systems that reject an attribute that cannot be proven to significantly influence accuracy. He states that pessimistic biases arise because criteria that are important in hypothesis testing (e.g., 95% confidence level) are naively translated into criteria that are used for evaluating "hypothesis plausibility" in learning systems. These types of criteria often prove to be overly exclusive, and their usage often results in the omission of significant attributes from models. In contrast to the pessimistic bias, Fisher proposes an optimistic bias. This bias states that an attribute is deemed relevant unless it is demonstrably non-influential to system accuracy. Fisher's research indicates that the choice between the optimistic and pessimistic biases in machine learning systems depends upon factors such as training sample size, the amount of noise in the training signal, and the distribution of outcomes across categories.

Training Sample Size. Fisher's findings show that optimism, a low confidence level for excluding attributes, is preferable when there is a smaller training sample. However, after the training sample reaches sufficient size, there is little difference in performances across confidence levels. Thus, in larger training samples, the choice between confidence levels for excluding attributes is not an important factor.

Training Signal Noise. Noise is variance in the training data that cannot be modeled by the learning algorithm. Fisher's research shows that optimism achieves better results in training samples with low noise. However, this preference declines, and can be reversed, as the training sample's size and noise increase. Thus, in large training samples with high noise, a more pessimistic bias may yield more accurate results.

Outcome Distribution. A relatively even distribution of prior probabilities for outcomes may be of more assistance in determining discrimination rules than a

distribution that is uneven. This is because an uneven distribution of prior probabilities creates a pre-existing bias in favor of one of the outcomes.

Fisher found that a pessimistic bias was detrimental to performance in a training sample where cases were evenly distributed across two outcomes. This was particularly true when there was a small training sample. Conversely, he found that the choice of confidence level had little effect on performance in a training sample where 91% of the observations were in one category. Thus, when there is a strong pre-existing bias caused by an uneven distribution of outcomes, neither pessimism nor optimism is preferred. However, when there is no pre-existing bias, optimism is preferred.

Attribute Order

Fryback demonstrated the importance of considering the sequence in which attributes are selected from a conditionally dependent training sample for inclusion in proper Bayes models.¹³ Although Fryback chose to include the most interdependent attributes first, it is more common to use some measure of the expected change in model accuracy as a guide in ordering attributes. Measures of expected attribute impact upon model accuracy that have been used to order model attributes include information gain, error reduction, and relative informativeness.^{16,19,20} Attribute order has no significance in the simple Bayes model where there is no reduction in the data set during computations. However, it can have significance in the proper Bayes model with conditionally dependent attributes where data-set reductions during computation may prevent some of the attributes from being included in the final model.

Model-factor Summary

Model factors have less potential for increasing the simple Bayes model's accuracy since this model does not partition its training sample during computation and, except in small samples, will not need to determine whether enough training sample cases are available to justify including an attribute in a model. The benefits model factors offer for the proper Bayes model are limited to specific situations. When model factors are used in situations other than these, they can decrease the accuracies of both Bayes models.

Study Methodology

This study investigated the impacts of variations in training sample sizes, numbers of attributes, and attribute exclusion criteria upon the accuracies of the simple and proper Bayes models. It was organized

Table 2 ■

Attributes Used to Predict In-hospital Mortality

Age	Mean arterial pressure
Arterial blood pH	Oxygenation
Chronic health	Potassium
Coma score	Respiratory rate
Creatinine	Sodium
Emergency admission	Surgical treatment
Heart rate	Temperature
Hematocrit	White blood cell count

as a controlled experiment, using a factorial design, in which these four attributes were manipulated while the remaining situational and model factors described above were controlled. The average areas under ROC curves were used to evaluate the discriminatory abilities of the models. The discriminatory abilities of the different models were compared by using pairwise *t* tests in multiple comparisons with a 0.05 level of significance.

Myocardial Infarction Database

All data used in this study were collected under a grant from the Health Care Finance Administration.²¹ Sixteen questions in that database correspond to the factors used in the APACHE II system to predict patient outcomes. These 16 questions were used as the attributes in our models. Another question in that database describes patient discharge status (coded as dead or alive). This question was used as the outcome that the models in this study predicted.²² The 16 model attributes are listed in Table 2. There were 1,139 cases in the myocardial infarction database that met the APACHE II inclusion criteria and were used in this study.

Training and Test Samples

Previous studies have compared the performances of different inductive learning systems.^{18,23} The performance criterion most frequently chosen is a system's ability to accurately classify cases that it has not previously seen. In our experiment, we randomly selected 339 cases as our test sample and 800 cases as our initial training sample. We then randomly split our 800 training cases into two other training samples of 100 and 400 cases. These training sample sizes were chosen because they were respectively below, within, and above the 200-to-500-case training sample range that Chard used to measure the simple Bayes model's performance.¹⁵ Chard reported that there were not enough cases for accurate predictions below this range and that the simple Bayes model's accuracy did not improve above this range. Although Chard's study differs from ours in many important

respects, it is nonetheless one of the few previous studies that have attempted to quantify the number of training sample cases required to accurately estimate the simple Bayes model's accuracy. For this reason, we used it as a benchmark in our study's selection of training sample sizes.

As an alternative design, we could have selected multiple random samples for each of our three training sample sizes. We did not choose this approach because our comparisons of average attribute values between the three training samples showed no significant difference for all 16 attributes between the 800- and 400-case samples and no significant difference for 15 of 16 attributes between the 800- and 100-case samples. Coma score was the only attribute that was different, and this difference occurred because there was no case with an extreme coma score in the 100-case training sample. Thus, although there were some differences between the three training samples, we did not deem these to be significant enough that multiple random samples were required.

Number of Attributes

We also selected attribute sets separately from our selection of training samples. Using a randomized procedure, we chose sets of four, eight, and 12 attributes from the 16 attributes in the APACHE II data set. Those attribute groupings were chosen because they were respectively below, near, and above the seven attributes used in Chard's study.¹⁵

We repeated the above procedure 30 times for each attribute-set size and created a total of 90 different attribute sets for use in this study. This design allowed us to manipulate the number of attributes while controlling the conditional dependence between attributes.

Next, we randomly assigned the attribute sets to the previously created training and test samples. This created ten occurrences for each of the nine training sample-size and attributes-set-size combinations. Thus, there were ten sample-size-attribute-set occurrences for each combination from 100 training cases and four attributes through 800 training cases and 12 attributes. This gave a total of 90 sample-size-attribute-set combinations.

Bayesian Expert System Shell (BESS) Models

In a previous study, we developed a system that selects its attributes according to their information value and is able to model zero through $n - 1$ orders of conditional dependence in Bayesian models.¹⁶ This system also may optionally use sample theory as a

means of excluding attributes from a model. This exclusion is effected by using the binomial distribution to test whether an attribute's likelihood ratio is statistically different from unity. If it is not statistically different from one, the attribute has no information value and is excluded. Kraemer and Thiemann provide tables that contain the number of training sample cases that are required to detect these differences for a particular likelihood ratio value.²⁴ Thus, by varying two of BESS's parameters—orders of conditional dependence managed by the model (zero for simple Bayes and $n - 1$ for proper Bayes) and use of sample theory to exclude attributes (yes for pessimism and no for optimism)—we were able to simulate four Bayes-model and attribute-selection scenarios. These were: 1) simple Bayes model with an optimistic attribute-exclusion criterion, 2) simple Bayes model with a pessimistic attribute-exclusion criterion, 3) proper Bayes model with an optimistic attribute-exclusion criterion, and 4) proper Bayes model with a pessimistic attribute-exclusion criterion. Although pessimistic exclusion is most often associated with the proper Bayes model, we decided to include scenario 2 since we believed that a pessimistic exclusion criterion could impact the simple Bayes model's accuracy when there was either a small training sample or a rare attribute value.

The last model factor, attribute order, was controlled by predetermining the order in which attributes would be entered into all of the Bayesian models. This procedure selected attributes whose values represented the most extreme likelihood ratios first. This ordering was chosen because the potentially most informative attributes are selected first. Since attribute order is not important for simple Bayes models, the use of a procedure that optimizes the entry of attributes in a model would affect the accuracy of the proper Bayes model only.

Outcome Prediction

We used BESS to calculate test-sample probabilities of patient survival under the four Bayes-model and attribute-selection scenarios for each of the 90 sample-size-attribute-set combinations. Using the calculated probabilities of patient survival, we then computed the area under the ROC curve for each of the 360 resulting combinations.^{25,26,27} This gave us ten measurements for each of the 36 distinct combinations of training sample size (100, 400, and 800 cases), number of model attributes (4, 8, and 12 attributes), Bayesian model (simple or proper), and attribute-exclusion criterion (sample theory or no exclusion). Last, we averaged the ROC measurements

across subsets of these data to create our outcome prediction measurement, average ROC area.

Results

SAS²⁸ statistical software was used to perform an analysis of variance (ANOVA). Table 3 shows the main and interaction effects that were identified in our experimental design. Higher-order interaction effects, although tested, are excluded from this table as they were not significant. The overall F test is computed by dividing the mean square for the model by the mean square for error. This test tells how well the model accounts for the dependent variable's (average ROC area) behavior. The F test is a test of the null hypothesis that all parameters except the intercept are zero. The significance probability associated with this statistic ($Pr > F$) is 0.0001, and this indicates that the experimental design accounts for a significant amount of the variability in average ROC area. The R-square statistic measures how much variation in the dependent variable can be accounted for by the model. The R-square value for this ANOVA is 57.586%.

Main Effects

All four main-effect variables in Table 3 have F tests that are significant at the 0.0001 level. This indicates that average ROC areas are not equal for different values of these variables and that the average ROC area can be expected to change as each of the main-effect variables individually changes its values.

Paired-interaction Effects

Only two interaction effects in our design were significant at or below the 0.05 level. These were training sample size with exclusion criterion and number of attributes with Bayes model. Two additional pairwise interactions that are cited in the literature (number of attributes with exclusion criterion and training sample size with Bayes model) did not have significant interaction effects. Table 4 gives the average ROC areas for these four pairwise interaction effects. Each cell in Table 4 is the average ROC area for one sixth of the sample points in this study. Thus, the upper left corner cell shows the average ROC area (0.73388) for the 100-case sample using the optimistic exclusion criterion. These measurements include all attribute set sizes (4, 8, and 12) and both Bayes models (simple and proper). The next cell in the first row shows the average ROC area for the 100-case sample using pessimistic exclusion (0.63620). The remaining cells in this row describe the 95% confidence interval for the difference between the optimistic and pessi-

Table 3 ■

Average ROC Area Analysis of Variance

$Pr > F = 0.0001$

R-square = 0.575860

ROC mean = .71378564

Root MSE = 0.052267

Effect	Pr > F
Main	
Training sample size	0.0001
Number of attributes	0.0001
Bayes model	0.0001
Exclusion criteria	0.0001
Interaction	
Training sample size · number of attributes	0.1476
Training sample size · Bayes model	0.3866
Training sample size · exclusion criteria	0.0001
Number of attributes · Bayes model	0.0073
Number of attributes · exclusion criteria	0.6408
Bayes model · exclusion criteria	0.6617

mistic means (-0.0977). The Significant Difference column denotes whether the difference in means is significant at the 0.05 level.

Training Sample Size with Exclusion Criterion. Table 4 shows that under optimistic exclusion (no attribute excluded) there is little change in the average ROC area (less than 1%) as the number of cases increases from 100 to 800. However, there is a significant change in the average ROC area (over 9%) as the number of cases increases under pessimistic exclusion (sample theory is used to exclude attributes). The average ROC areas for 100 and 400 cases with optimistic exclusion are significantly different from the corresponding average ROC areas with pessimistic exclusion. However, there is no difference between optimism and pessimism with 800 cases. Thus, with a small or intermediate number of cases (100 or 400), optimism is preferred to pessimism, and with a large number of cases (800), there is no difference in accuracy between the optimistic and pessimistic exclusion criteria.

Figure 1 extends this analysis by separating exclusion criteria by type of Bayesian model. There are lines in this graph for: 1) proper Bayes model with no sample-theory exclusion, 2) proper Bayes model with sample-theory exclusion, 3) simple Bayes model with no sample-theory exclusion, and 4) simple Bayes with sample-theory exclusion. This graph shows that, regardless of whether the simple or proper Bayes model is used, optimism (no sample-theory exclusion) is

Figure 1 Sample size and ROC area. Star indicates simple Bayes with optimistic exclusion; hollow square, simple Bayes with pessimistic exclusion; shaded square, proper Bayes with optimistic exclusion; diamond, proper Bayes with pessimistic exclusion.

preferable to pessimism (sample-theory exclusion). And, as the number of training cases increases, the four lines begin to converge until at 800 cases there is little difference between three of the lines. Only the proper Bayes model with sample-theory exclusion lags behind the other models.

Number of Attributes with Bayes Model. With regard to the second significant interaction effect, number of attributes with Bayes model, Table 4 shows that the simple and the proper Bayes models both increase

their accuracies as the number of attributes increases. However, the simple Bayes model increases its accuracy at a greater rate than does the proper Bayes model. At four attributes, there is no difference between the models, but at 12 attributes the simple Bayes model is over 4% more accurate than the proper Bayes model. There is no significant difference in average ROC areas for the two Bayes models with four attributes. However, at both eight and 12 attributes, there are significant differences. Thus, with a small number of attributes (four), there is no differ-

Table 4 ■

Average ROC Area by Interaction Effects

Exclusion Criteria						
Situational Factors	Optimistic Mean	Pessimistic Mean	95% Confidence Interval for Difference in Means			Significant Difference*
			Lower Limit	Difference in Means	Upper Limit	
100 cases	0.73388	0.63620	− 0.1217	− 0.0977	− 0.0737	Y
400 cases	0.74238	0.70041	− 0.0660	− 0.0420	− 0.0180	Y
800 cases	0.74152	0.72832	− 0.0108	0.0132	0.0372	N
4 attributes	0.68934	0.63104	0.0368	0.0583	0.0798	Y
8 attributes	0.74805	0.70114	0.0254	0.0469	0.0684	Y
12 attributes	0.78039	0.73275	0.0261	0.0476	0.0692	Y
Bayes Models						
	Simple Bayes Mean	Proper Bayes Mean	95% Confidence Interval for Difference in Means			
			Lower Limit	Difference in Means	Upper Limit	
100 cases	0.70027	0.66982	0.0044	0.0305	0.0566	Y
400 cases	0.72801	0.71477	− 0.0129	0.0132	0.0393	N
800 cases	0.74894	0.72090	0.0019	0.0280	0.0541	Y
4 attributes	0.66125	0.65913	− 0.0207	0.0021	0.0249	N
8 attributes	0.73702	0.71218	0.0020	0.0248	0.0477	Y
12 attributes	0.77895	0.73418	0.0220	0.0448	0.0676	Y

*Y = yes; N = no.

Figure 2 Number of attributes and ROC area. Star indicates simple Bayes with optimistic exclusion; hollow square, simple Bayes with pessimistic exclusion; shaded square, proper Bayes with optimistic exclusion; diamond, proper Bayes with pessimistic exclusion.

ence in accuracy between the two Bayes models, and with a moderate to large number of attributes (eight or 12), the simple Bayes model is preferred.

Figure 2 extends this analysis by separating types of Bayes models by the two exclusion-criterion options. This creates in Figure 2 the same four lines that appeared in Figure 1. This graph shows that the major division between the models at four attributes is the choice of exclusion criterion and not the type of Bayes model. At eight and 12 attributes, all four lines are increasing. However, the simple Bayes models, with

and without sample-theory exclusion, increase in accuracy at a greater rate than the proper Bayes models.

Nonsignificant Interaction Effects. The two pairwise interaction effects that are mentioned in the literature and that were not significant in this design each present a special case. With regard to number of attributes with an exclusion criterion, it is always better to use optimistic exclusion. Optimism at four, eight, and 12 attributes consistently produces more accurate estimates than pessimism. This relationship is also depicted in Figure 2. Thus, with a small, interme-

Table 5. ■

Average ROC Areas with Pessimistic and Optimistic Exclusions

Training Sample Size	Number of Attributes	Simple Bayes Mean	Proper Bayes Mean	95% Confidence Interval for Difference in Means			Significant Difference*
				Lower Limit	Difference in Means	Upper Limit	
Pessimistic							
100 cases	4 attributes	0.56091	0.56013	−0.0452	0.0008	0.0468	N
	8 attributes	0.67927	0.64946	−0.0162	0.0298	0.0758	N
	12 attributes	0.71208	0.65536	0.0107	0.0567	0.1027	Y
400 cases	4 attributes	0.64964	0.65613	−0.0525	−0.0065	0.0395	N
	8 attributes	0.70857	0.69682	−0.0342	0.0117	0.0577	N
	12 attributes	0.76013	0.73117	−0.0170	0.0290	0.0750	N
800 cases	4 attributes	0.68575	0.67370	−0.0339	0.0121	0.0580	N
	8 attributes	0.75768	0.71507	−0.0034	0.0426	0.0886	N
	12 attributes	0.79924	0.73850	0.0147	0.0607	0.1067	Y
Optimistic							
100 cases	4 attributes	0.68161	0.67926	−0.0436	0.0023	0.0483	N
	8 attributes	0.76841	0.74040	−0.0180	0.0280	0.0740	N
	12 attributes	0.79933	0.73428	0.0191	0.0650	0.1110	Y
400 cases	4 attributes	0.69159	0.69251	−0.0469	−0.0009	0.0451	N
	8 attributes	0.75523	0.74309	−0.0338	0.0121	0.0581	N
	12 attributes	0.80293	0.76891	−0.0120	0.0340	0.0800	N
800 cases	4 attributes	0.69800	0.69304	−0.0410	0.0050	0.0509	N
	8 attributes	0.75298	0.72821	−0.0212	0.0248	0.0708	N
	12 attributes	0.80002	0.77689	−0.0229	0.0231	0.0691	N

*Y = yes; N = no.

diate, or large number of attributes, optimism, with either Bayes model, is always preferred to pessimism.

With regard to training sample size with Bayesian models, the situation is more complex. Table 4 shows that with a small number of cases (100) and with a large number of cases (800), the simple Bayes model is more accurate. However, with an intermediate number of cases (400), there is no difference in accuracy between the models. This relationship between training sample size and Bayes models can be attributed to the fact that the proper Bayes model begins to show signs of peaking between 400 and 800 cases while the simple Bayes model does not.

Complex Interaction Effect

One higher-order interaction effect, training sample size with number of attributes and Bayes model, is often cited in the literature. The assumption is that the proper Bayes model will outperform the simple Bayes model when there is a large training sample size with a large number of attributes. However, this interaction was not significant in our design. Table 5 summarizes this relationship for pessimistic and optimistic exclusions. While the simple Bayes model's accuracy is consistently equal to or greater than that of the proper Bayes model, the differences in accuracy are significant only in the 12-attribute models with 100 or 800 cases for pessimistic exclusion and in the 12-attribute model with 100 cases for optimistic exclusion.

Discussion

Situational Factors

Prior research has shown that when conditional dependence exists, the simple Bayes model may yield results that are equal to or better than those of the proper Bayes model, and that increasing the training sample size will have a greater impact on the accuracy of the proper Bayes model than on that of the simple Bayes model.^{8,13} Other researchers have proposed that with large training-sample sizes, the proper Bayes model will produce more accurate results than the simple Bayes model.¹¹ Our results confirm that the simple Bayes model can be more accurate than the proper Bayes model when there is conditional dependence and that the proper Bayes model becomes more accurate as training samples increase in size. However, the accuracy of the proper Bayes model never exceeded that of the simple Bayes model. In fact, across all attribute-set sizes and for both pessimistic and optimistic exclusions, the proper Bayes model seemed to begin peaking between 400 and 800

cases. Although previous researchers have reported that models that partition the training sample begin to peak when the ratio of the training sample size to the number of attributes is small, they have not reported that these models begin to peak as additional training-sample cases are added.¹⁴

The fact that our simple Bayes models did not become more accurate as the training sample increased in size seems to conflict with the results of Chard.¹⁵ This earlier study concluded that a minimum of 200 cases is required for the simple Bayes model and that there was a maximum of approximately 500 cases above which increasing the training-sample size would produce no further improvement in the model's accuracy. In contrast, our model achieved a point of maximum accuracy at or below 100 training cases. Despite the seeming contradiction, there are important differences between Chard's study and ours. Chard was concerned with estimating eight outcomes from seven attributes. The prior probabilities for his outcomes ranged from 2% to 30%. Our study estimated one outcome from four, eight, and 12 attributes and our prior probability was 85%. Thus, with a larger prior probability and fewer outcomes to estimate, we could be expected to achieve more accurate results with a smaller sample.

Several researchers have reported smaller differences in accuracy between the simple and proper Bayes models with smaller numbers of attributes and larger differences in accuracy with larger attribute sets.^{13,14} These authors have also concluded that the proper Bayes model should be used only with a large training sample where the conditional dependencies in the proper Bayes model can reliably be estimated. Our study found that with a small number of attributes, there was little difference between the performances of the simple and proper Bayes models. However, as more attributes were added, it was the simple, rather than the proper, Bayes model that had the greater improvement in accuracy.

Our results from simultaneously manipulating training sample size and number of attributes were also different than anticipated. The one situation where the proper Bayes model is generally agreed to outperform the simple Bayes model is where there are conditional dependence in the data, a large training sample, and a large number of attributes. In contrast, our results showed that the proper Bayes model never outperformed the simple Bayes model.

Manipulating the number of attributes in a simple Bayes model appears to have a far greater impact on model accuracy than is achieved by increasing the

training sample size. In fact, it may be more economical in many studies to increase a model's accuracy by increasing the number of attributes that are collected rather than by increasing the training sample size. Again, this relationship should be investigated further.

Model Factors

Fisher observed that an optimistic bias is preferred when there is a small training sample and that neither bias is preferred with a large training sample.¹⁸ Our results support this observation. With a small or intermediate number of cases (100 or 400), optimistic exclusion produced more accurate estimates. With a larger number of cases (800), there was no significant difference in accuracy with either pessimistic or optimistic exclusion. We extended Fisher's work by varying the number of attributes in our models. We found that optimistic exclusion consistently outperformed pessimistic exclusion for all attribute sets (small, intermediate, and large).

Conclusion

Our results support the position that managing conditional dependence may not be as important a problem as some researchers have assumed and that the simple Bayes model may be fully capable of managing conditional dependence in data. In our study, the simple Bayes model consistently equalled or outperformed the proper Bayes model, even in the one situation (large training sample and large number of attributes with conditionally dependent data) where the proper Bayes model had been assumed to excel.

The fact that the proper Bayes model did not outperform the simple Bayes model as the training sample size increased contradicts assumptions that are prevalent in the literature. This finding brings into question whether the distinction between the appropriate usage of simple versus proper Bayes models can be based upon training sample sizes.^{13,14} It also questions whether the greater accuracy of the simple Bayes model versus the proper Bayes model in published studies can be attributed to a lack of adequate sample size.¹¹ What is evident is that more research is needed to determine all of the situational factors that are related to the accuracy of proper Bayes models.

Two of the weaknesses of this study were that only one medical domain (myocardial infarction patients measured using APACHE II data) was included and that no measurement of conditional dependence was proposed. It is quite possible that many of the results

of this study are related to the data sets that were used and cannot be generalized to other domains. In order to test the generalizability of these results, other similar studies would need to be performed in other medical domains and their results compared with ours. With regard to conditional dependence, no attempt was made to measure the amount of conditional dependence in our data sets and to vary this as a situational factor in our experimental design. Clearly, conditional dependence and its effect on model accuracy is a complex concept that needs to be addressed in future studies.

Previous researchers using Bayesian belief networks have made significant contributions to the understanding of conditional dependence and its management in Bayesian models. Unfortunately, probabilistic inference using belief networks has been shown to be NP-hard.²⁹ Consequently, approximate and special-case algorithms for belief-net inference are required.^{30,31} Another difficulty with belief networks is the fact that they all require some descriptive domain knowledge in their computation.^{32,33} In contrast, our research has attempted to apply Bayesian models without assuming prior domain knowledge. In this regard, it has more affinity with machine induction systems and neural networks than it does with Bayesian belief networks.³⁴

While this study did not discover those situational factors that would allow the proper Bayes model to outperform the simple Bayes model, it did eliminate several candidates (sample size, number of attributes, and exclusion criteria). The study also demonstrated that increasing sample size might not be the best way of handling problems of this type. Instead, other approaches that would better manage the data that are available should be investigated. Some of these other approaches include: better models for ordering model attributes, exclusion criteria between optimism and pessimism, and using a less exact definition of conditional dependence and managing some, but not all, of the dependencies.

References ■

1. Gustafson DH, Cats-Baril WL, Alemi F. Systems to Support Health Policy Analysis: Theory, Models, and Uses. Ann Arbor, MI: Health Administration Press, 1992.
2. Gustafson DH, Kestly JJ, Greist JH, Jensln NM. Initial evaluation of a subjective Bayesian diagnostic system. *Health Serv Res.* 1971;6:204.
3. Norusis MJ, Jacquez JA. Diagnosis. I. Symptom nonindependence in mathematical models for diagnosis. *Comput Biomed Res.* 1975;8:156-72.
4. Seroussi B, the ARC and AURC Cooperative Group. Computer-aided diagnosis of acute abdominal pain when taking

- into account interactions. *Meth Inform Med.* 1986;25:194–8.
5. Wu Y-C. Artificial neural networks for knowledge representation: analysis of problem structure based on the Bayes' theorem and conditional nonindependence. PhD Dissertation. Madison, WI: University of Wisconsin, 1992.
6. de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of abdominal pain. *Br Med J.* 1972;II:9–13.
7. Spicer CC. Test reduction: II—Bayes's theorem and the evaluation of tests. *Br Med J.* 1980;30:592–4.
8. Russek E, Kronmal RA, Fisher LD. The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis. *Comput Biomed Res.* 1983;16: 537–52.
9. Weintraub WS, Madeira SW, Bodenheimer MM, et al. Critical analysis of the application of Bayes' theorem to sequential testing in the noninvasive diagnosis of coronary artery disease. *Am J Cardiol.* 1984;54:43–9.
10. Bayes T. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions* 1763;3:317–418. Reproduced in *Two Papers by Bayes*, ed. W. E. Deming. New York: Hafner, 1963.
11. Gammerman A, Thatcher AR. Bayesian diagnostic probabilities without assuming independence of symptoms. *Meth Inform Med.* 1992;30:15–22.
12. Chard T. The effect of dependence on the performance of Bayes' theorem: an evaluation using a computer simulation. *Comput Meth Prog Biomed.* 1989;29:15–9.
13. Fryback DG. Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Comput Biomed Res.* 1978; 11:423–34.
14. Ohmann C, Qin Yang, Kunneke M, Stoltzing H, Thon K, Lorenz W. Bayes theorem and conditional dependence of symptoms: different models applied to data of upper gastrointestinal bleeding. *Meth Inform Med.* 1988;27:73–83.
15. Chard T. Self-learning for a Bayesian knowledge base: how long does it take for the machine to educate itself? *Meth Inform Med.* 1987;26:185–8.
16. Alemi F, Bhatt P, Eisenstein E, Fadlalla A, Stephens R, Butts J. Torturing data until they confess: a self learning Bayesian expert system (B.E.St.). Working Paper of the Collaborative Care Project, Cleveland, OH, 1992.
17. Quinlan R. Induction of decision trees. *Machine Learning.* 1986;1:81–106.
18. Fisher D. Pessimistic and Optimistic Induction, Technical Report CS-92-12, Department of Computer Science, Vanderbilt University, Nashville, TN, 1992.
19. Bigongiari LR, Preston DF, Cook L, Dwyer SJ, Fritz S, Fryback DG, Thornbury JR. Uncertainty/information as measure of various urographic parameters: an information theory model of diagnosis of renal masses. *Invest Radiol.* 1981;16:77–81.
20. Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.
21. Alemi F, Rice J, Hankins R. Predicting in-hospital survival of myocardial infarction. *Med Care.* 1990;28(9):762–76.
22. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med.* 1985;13(10):818–29.
23. Kononenko I, Bratko I. Information-based evaluation criterion for classifier's performance. *Machine Learning.* 1991;6:67–80.
24. Kraemer JC, Thiemann S. How many subjects: statistical power analysis and research. Newbury Park, CA: Sage, 1987.
25. Hanley J, McNeil B. The meaning and use of the area under a receiver operating curve. *Diagnost Radiol.* 1982;143:29–36.
26. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same data. *Radiology.* 1983;148:839–43.
27. McNeil BJ, Hanley J. Statistical approaches to the analysis of receiver operating characteristic curves. *Med Decis Making.* 1984;4:137–50.
28. SAS Institute Inc. SAS/STAT User's Guide, Release 6.03 Edition. Cary, NC: SAS Institute Inc., 1988.
29. Cooper GF. The computational complexity of probabilistic inference using Bayesian belief networks. *Artif Intell.* 1990;42:393–405.
30. Herskovits EH, Cooper GF. Algorithms for Bayesian belief-network precomputation. *Meth Inform Med.* 1991;30:81–9.
31. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. *J R Statist Soc B.* 1988;50:157–224.
32. Harris N, Spiegelhalter DJ, Bull K, Franklin CG. Criticizing conditional probabilities in belief networks. *Proc Fourteenth Annu Symp Computer Applications in Medical Care.* 1990:805–9.
33. Herskovits E. Computer-based probabilistic-network construction. PhD dissertation, Stanford University, Stanford, CA, 1991.
34. Weiss SM, Kulikowski CA. Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Mateo, CA: Morgan Kaufmann, 1991.