Evaluation of probabilistic consultation systems by simulation

Svein Nordbotten
Department of Information Science
University of Bergen
N-5000 Bergen
NORWAY

This paper reports on an evaluation of alternative inference models applied in probabilistic consultation systems by simulation. In the first section some general remarks about consultation systems are made in order to set the specific system discussed in a wider context. Section 2 gives a description of the different inference models tested. The next section specifies simulation experiments which aim at an evaluation of the inference algorithms. Section 4 outlines the actual experiments and in Section 5 the results of the experiments are presented. The main conclusions supported by the experiments are: the rule value model gives more successful predictions than simpler inference models, that a model applying the complete Bayes' formula performs better than one which instead makes use of complementary probabilities in the denominators, and that the estimates of prediction probabilities in models ignoring the statistical dependence among symptoms tend to be highly biased and cannot be expected to be useful reliability indicators.

Keywords: Expert system, Inference model, Simulation, System evaluation

1. Consultation systems
1.1 Types of consultation systems

A consultation system, also called expert or knowledge-based system, is a computer-based system for assisting a non-professional in solving problems which usually require the assistance of a specialist or expert. The range of a consultation system is usually limited to a particular subject matter field. As a result of increased funding for research in artificial intelligence, consultation systems have proliferated in the software market in the last five to seven years and a great variety of systems and tools for different purposes have emerged.

One basic property of consultation systems is the separation of the inference program from the knowledge base to make the knowledge base easily accessible for modifications. We distinguish between complete consultation systems for assisting the end users to solve their problems within a particular domain, and empty systems or shells, which are tools for building such complete systems if a knowledge base can be acquired and added to the shell.

The way knowledge is represented in a consultation system is another basic aspect of the system. A number of different schemes have been suggested. In most, however, the IF-THEN rule representation is a basic component. The experiments discussed in the present paper are based on a knowledge base with such rules.

Most consultation systems try to represent informal knowledge gained through experience. In contrast to systems working with exact associations or with knowledge which can be satisfactorily approximated by exact associations, uncertainty is an important aspect of the knowledge in the system which we investigate. One approach to uncertainty is to apply probability theory, other means are fuzzy logic and certainty indicators. In this paper we focus on the application of probability theory.
Consultation systems may also be distinguished by the type of problems they can advise on. A very common problem is the classification problem. In a classification problem, an object, real or abstract, is to be classified by an attribute which is not directly observable. Instead, one or more attributes related to the classification attribute can be observed or investigated and form the basis for the classification. Classical examples are classification of animals by family based on their specific properties, classification of patients by disease based on their symptoms, etc.

The inference strategy used by a consultation system for searching the knowledge to be used for task solving, is an important feature of the system. Several approaches can be distinguished as backward chaining, forward chaining and sideways chaining or rule value approach. While the two first approaches are based on symbol string matching either when the knowledge base is created or during the processing of a task, the rule value approach is based on evaluation of the importance of each rule during search for a solution.

1.2 Stochastic classification problems

In the subsequent discussion, we shall use the following terminology. An incident is the object of the current task. Each incident is described by some unknown, but observable symptoms and an unobservable diagnosis. The task is to predict the diagnosis of the incident, or to classify the incident by diagnosis.

Consider a set D=\{d[1]..d[N]\} of diagnoses, a set S=\{s[l]..s[M]\} of symptoms, a set E=\{e[1]..e[J]\} of all subsets of S and the product set C=E*D with the elements d[k]=\{e[j]\}d[i]. The C set represents all possible occurrences of symptoms and diagnoses. We introduce a set P with the elements p[k]=\{d[j]\}d[i]. The elements of P satisfy:

\[ P(d[i],d[j])=P(e[j],d[i])/P(d[i]) \quad \text{for all } i=1..N, \quad j=1..M. \]

By means of Bayes' Theorem, the probability for a diagnosis, d[i], when a symptom set, e[j], is known to occur in the incident, is:

\[ P(d[i]|e[j])=P(e[j]|d[i])P(d[i]) \quad \text{for all } i=1..N, \quad j=1..J. \]

This is our general model. We shall in the subsequent sections make further specifications and modifications.

1.3 Classification tasks

A set of incidents of elements of C are assumed to be produced by random according to the probabilities of P. The d-element of the incident cannot be observed by the investigator, the e-element can be investigated subject to given conditions. The task of probabilistic classification can be considered as a search for an element of C which has an e-element which matches the detected symptoms of the observed incident and produces a satisfactory prediction of the d-element. The search has to be carried out by investiga-
when the symptom investigated occurs, or
(2.2) \( P(d|\neg s) = \frac{(1-P(s)P(d)|s)}{(1-P(s)P(d)|s)+(1-P(s)\neg P(d)|\neg s))} \)
when the symptom does not occur. The computed conditional probabilities become the new probabilities of the diagnoses when the symptom has been investigated. This formula requires less computations than the complete formula (1.4), but requires on the other hand the availability of the probabilities \( P(s|d) \). The sets of rules needed in the knowledge base by this model are:
(2.3) IF \( d \) THEN \( P(s|d) \) is the probability for \( s \).
and
(2.4) IF NOT \( d \) THEN \( P(s|\neg d) \) is probability for \( s \).

The search for symptoms proceeds in steps. At each step the symptom is selected which will have the strongest influence on the probabilities for prediction. The expression:
(2.5) \( RV(s) = \sum d \text{ abs}(P(d) - P(d|\neg s)) \)

is computed for each symptom not yet investigated. The rule for selecting the next symptom to be investigated is then:
(2.6) IF \( RV(s') \) is max. for all \( j \) not yet investigated

THEN select \( s' \) as the next symptom to investigate.

Naylor indicates that the rule value may alternatively be computed in other ways instead of using the absolute value. Investigation of a selected symptom may result in either concluding that the symptom does occur or that it does not occur. For each step the conditional probability of each diagnosis is recalculated according to the rule (2.1) or (2.2) depending on the outcome of the investigation of symptom \( s' \).

The criteria used for stopping the search for more symptoms and selecting a diagnosis, requires that the prediction must be the most probable, even if a full investigation was made of all the uninvestigated symptoms, and the outcome of this investigation turned out in favor of other predictions:
(2.7) IF
1. \( P(d'|s') \) is max. in \( D \)
AND
\( P(d'|s') > P(d|s'|s') \) \( E-s[f'] \)
AND
\( P(d'|s') = \text{max. in } D \)
OR
2. [all elements of \( S \) are investigated]

THEN

decide that \( d' \) is the prediction.

where \( i' \) and \( j' \) have the same meaning as above, and \( E-s[f'] \) is the complementary subset to \( d' \) in \( E \).

There are several objections to this model. The main objection is that the model requires extensive computations and if not, a less complex model may be competitive.

From a statistical viewpoint, the objections are that in real life applications, the assumption about statistical independence is likely not to be valid. When Model 1 is applied to incidents in which there is statistical dependence among the symptoms, the probabilities of the predictions are expected to be overestimated. The use of the simpler Bayesian formula (2.1) or (2.2) assumes perfect knowledge of the second set of rules (2.4) which may be unrealistic and result in less precise estimates than the use of a complete formula, would have given. Repeated use of the formula without re-calculating the denominator, also raises serious questions.

2.3 Inference Model 2

The inference model 2 represents a more naive and computationally simpler type of inference. It assumes that the search for a symptom to be investigated should at each stage start by the diagnosis having the highest probability. The symptom selected for investigation should then be the symptom with the highest conditional probability given that the most likely diagnosis is true. If during the investigation another diagnosis becomes the most probable, the symptoms of this diagnosis are pursued.

The rule for selecting a symptom for investigation is:
(2.8) IF
\( P(d'|s') = \text{max. for } d \) in \( D \)
AND
\( P(s'|d') = \text{max. for } s \) in \( S-u[f'] \)

THEN
investigate \( s' \).

where \( i' \) is the index of the most probable diagnosis, \( s[f'] \) the symptom set so far detected, \( u[f'] \) the symptom set investigated, \( s[k'] \) the symptom to be investigated and \( S-u[f'] \) is the set of all symptoms not yet investigated.

The investigation of symptoms is carried out until all symptoms related to the currently most probable diagnosis have been investigated.

(2.9) IF
\( P(d'|s') = \text{max. for } d \) in \( D \)
AND
\( P(s'|d') = 0 \) for all \( s \) in \( S-u[f'] \)

THEN
decide that \( d' \) is the prediction.

This model thus ignores the possibilities that if all symptoms were investigated, some other diagnoses might be more probable.

For each step, the probabilities for each alternative diagnosis is recalculated according to (2.1) or (2.2) depending on the outcome of the symptom investigation.

This model of inference has a more intuitive appeal since it can be more easily understood and represents what most people believe they would do themselves. It also requires less computation that the Model 1. The question is how successful it is in making correct predictions. It should be expected that the number of incorrect predictions produced by this model will be less than produced by Model 1.

2.4 Inference Model 3

One way of reducing the computational requirements of Model 1 is to use the simpler stop rule used in Model 2. Model 3 is like Model 1 with the exception that it uses the decision (2.9) instead of the more complex and probably safer rule (2.7). It is expected that this model will require
less resources, but that it may also produce more failures than Model 1.

2.5 Inference Model 4

The three previous models all apply the simple Bayes’ formula in computing the probabilities. Model 4 is a further development of Model 2 and applies the complete Bayes’ formula as a basis for computing the probabilities:

\[
P(d[i] \mid s[j]) = \frac{P(s[j] \mid d[i]) \cdot P(d[i])}{\text{SUM}[i]P(s[j] \mid d[i])P(d[i])}
\]

or

\[
P(d[i] \mid \text{NOT } s[j]) = \frac{(1 - P(s[j] \mid d[i])) \cdot P(d[i])}{\text{SUM}[i] (1 - P(s[j] \mid d[i])) \cdot P(d[i])}
\]

to compute the conditional probabilities of the alternative diagnoses. It does not need the second set (2.4) of rules. Model 4 requires, however, more computations than Model 2 does, and it is expected that it will provide more precise estimates of the probabilities than Model 2 does.

2.6 Inference Model 5

Model 5 relaxes the assumption about independent symptoms. It applies the same rule value procedure as Model 1, but differs from that model in two respects. It stops when the rule (2.7) is satisfied, but it does not use the prediction as prescribed by this rule. It first recalculates the conditional probabilities for all diagnoses based on the frequencies for the symptom complexes \( d[j] \) detected:

\[
P(d[i] \mid \text{el}^*) = \frac{P(e[j] \mid d[i]) \cdot P(d[i])}{\text{SUM}[i]P(e[j] \mid d[i])P(d[i])}
\]

for \( i = 1 \ldots N \) and \( j \) is the index of the symptom complex detected. Since the recalculation only is carried out once on bases of the detected set \( e[j] \), we do not need a counterpart to the formula (2.2).

Then it looks for the diagnosis with the maximum probability which is then used as the prediction.

Model 5 requires an advanced knowledge base which, in addition to the rules of the knowledge base used by Model 1, must also include these probabilistic rules:

\[
\begin{align*}
\text{IF } d[i] \text{ THEN } P(e[j] \mid d[i]) &= \text{the probability for } e[j] \\
\text{and} \\
\text{IF NOT } d[i] \text{ THEN } P(e[j] \mid \text{NOT } d[i]) &= \text{is probability for } e[j].
\end{align*}
\]

Since the number of possible elements in \( E \) will increase very fast with increasing number of elements in \( S \), we can hope only to be able to acquire the most usual elements and rules.

Because of this, it should be expected that this model will not be able to identify all possible symptom complexes and produce predictions for all possible incidents. The incidents for which it will be able to provide estimates, may be assigned predictions different from those provided by Model 1. Because of the more consistent application of the conditional probabilities, the probability estimates for the incorrect predictions should be lower than those produced by Model 1.

3. Design of experiments

3.1 Objectives

The main objective of this paper is to investigate and test the following hypotheses by simulation experiments:

1) Hypothesis 1: Search, estimation and stop strategy of Model 1 does give a higher quality and efficiency than the less sophisticated strategy of Model 2.

2) Hypothesis 2: The complex stop rule of Model 1 yields a significantly higher quality or better efficiency than the simpler decision rule of Model 3 for selecting a prediction.

3) Hypothesis 3: Using the complete Bayes’ formula of Model 4 contributes significantly to higher quality and more efficient performance than the simpler formula of Model 2.

4) Hypothesis 4: There is a significant difference in the performances of Model 1 and Model 5 as to quality and efficiency on a set of incidents with symptoms which are statistically dependent.

5) Hypothesis 5: There is no significant difference between the estimates of the probabilities for the correctly and the incorrectly predicted diagnoses in the models studied.

In the statistical tests, we define the significant difference by a critical value based on a confidence coefficient 0.95 meaning that the probability of rejecting a correct hypothesis is less than 5%.

According to section 1.3, we assume that the quality of a prediction is expressed by a function \( q \) of the predicted and the true diagnosis with the value 1 if the two arguments are equal and 0 otherwise. The true diagnosis is known under our ideal experimental conditions, and we shall be able to observe the exact quality of the predictions. Under less ideal, but more realistic conditions, the true value will not be available. It is desirable that the quality can be indicated by the estimated probability of the prediction.

We assume that the cost of a prediction is expressed by a function \( q \) equal to the a weighted sum of the number of symptom complexes investigated, the number of incorrect predictions and the time consumed by the inference process. To describe the efficiency of the performance of a model we use a ratio between total quality and total cost for a given set of task solutions. We emphasize that it is not an objective of the present paper to study how well a consultation system performs compared with the performance of experts in some specific subject matter domain. For our purpose, we assume that the knowledge base is a perfect representation of the expert knowledge and that we want to study how well the inference models with their assumptions are able to utilize the knowledge base and produce correct conclusions.

3.2 The test bench

To accomplish the objectives, we need to establish a test bench for our 12 experiments. The main tools for carrying out our experiments are:

— a program system in which our inference models are implemented as alternative modules,
one knowledge base which represents typical expert knowledge and experience represented by lists of symptoms, diagnoses and stochastic rules of the type (2.3) and (2.4). The probabilities of (2.4) in acquired knowledge need not be consistent with (2.3).

a second knowledge base with the same content as the above, but with the probabilities of (2.4) adjusted so they are consistent with (2.3).

a third, advanced knowledge base which, in addition to the components of the first base, also includes rules of type (2.13) and (2.14),

one set of recorded incidents which represent tasks with statistically independent symptoms,

a second set of recorded incidents which represent tasks with statistically dependent symptoms,

a statistical program system which processes the outcomes of the consultation system simulations and computes necessary statistics,

a computer which can run the consultation system and represents the type of hardware we think should run a consultation system.

The program system must be able to process the cases presented according to the five models discussed above. It must also be able to process a set of incidents without interaction with the experimenter. The system must record for each incident the number of symptoms it investigated, specified by number of confirmative and non-confirmative observations and the predicted diagnosis with the estimated probability in a log together with the "true", for the system hidden diagnoses, such that they can later be analyzed.

The knowledge bases used may be borrowed from any subject field as long as they seem to have representative characteristics. They must contain representations of a set of symptoms, a set of conclusions with their associated a priori probabilities, and the sets of probability rules described above. To have a realistic size order, we will assume that a knowledge base must contain a set of symptom representations with more than 50 elements, a set of represented conclusions with more than 75 elements, and a set of statistical representations between at least 500 conditional probabilities for symptom given conclusion, and the same number of rules for conditional probabilities for symptoms given NOT conclusion for those models which require such rules. The advanced base should in addition also have at least $2^{330}$ conditional probabilities for symptom complexes.

Two test sets of incidents are required. Each incident within a test set must specify a conclusion and one or more occurring symptoms. The conclusion will be hidden for the consultation system, while the occurrence of symptoms can be investigated. One set should be a random sample of incidents from the domain which the knowledge bases represent. The symptoms of each incident of this set should be independent of the occurrence of each other, but consistent with the conditional probabilities for symptoms given the diagnosis of the incident:

\[(3.1) \ P(s[j] \mid s[k], d[i]) = P(s[j] \mid d[i]) \text{ for all } i,j \text{ and } k.\]

The second incident set required, is also a random sample from the knowledge base, but the symptoms of each incident in this set must be statistically dependent on each other, that is:

\[(3.2) \ P(s[j] \mid s[k], d[i]) \neq P(s[j] \mid d[i]).\]

for some symptoms, at the same time as the probabilities:

\[(3.3) \ P(s[j] \mid s[k], d[i]) \text{ and } P(s[j] \mid d[i])\]

of the knowledge base are maintained. This implies that:

\[(3.4) \ P(s[j] \mid \text{NOT } s[k], d[i])<>P(s[j] \mid d[i])\]

for some symptoms.

Each set must have a size which permits conclusion to be drawn at a high confidence level. Based on general statistical considerations, we decided that the size of the two sets should be 400 incidents.

The statistical system must be able to provide tabular summaries of each log set, compute the necessary statistical variables required by the relevant tests and indicate the statistical conclusions assuming given criteria.

The hardware which the experiments require, must have capacity, speed and functionality to run the program system within reasonable time limits.

### 3.3 Test variables and statistics

To evaluate the different alternatives, the logs taken must include the following measures, recorded for each processed incident:

- **Diagnosis:** the "true" diagnosis.
- **A priori probability:** the probability before any investigation.
- **Occurring symptoms:** number of occurring symptoms.
- **Prediction:** the diagnosis predicted.
- **Estimated probability:** the estimated probability for prediction.
- **Investigations:** number of symptoms investigated.
- **Detected symptoms:** number of occurring symptoms detected.
- **Time:** processing time.

In addition some test statistics must also be computed for each run to permit some statistical tests. Those include chi-square value for the differences between expected and true distributions by diagnosis, chi-square value for the differences between true and predicted distribution, and $F$ value for the difference between estimated probabilities for the group of correct and the group of incorrect predictions.

### 3.4 Test runs

The experiments will require the following runs:

**A. Runs on the incidents with independent symptoms**

Run 1:
In this run Model 1 is applied on the set of incidents with independent symptoms. This set is used for two reasons. First, the model assumes independence and second, the prediction is not less difficult when the symptoms of the
incidents are independent of each other. The output of this run is a set called Log 1.

Run 2: This run applies Model 2 with its simpler search process and stop rule for selecting a prediction. The output permits a comparison between the two models. The results of this run form Log 2.

Run 3: Run 3 is based on Model 3. Since this model contains the same inference process as that used in Model 1, but the stop rule of Model 2, the output permits an examination of the impact of the stop rules. The output of this run is referred to as Log 3.

Run 4: Model 4 differs from Model 2 by not using the second set of conditional probability rules, but instead the complete Bayes’ formula (2.10). The output, Log 4, together with Log 2 represents a basis for studying the effect of the complete Bayes’ formula.

Run 5: Model 5 with the advanced knowledge base is used in this run. This model assumes incidents with independent symptoms to provide more knowledge about the use of the complete Bayes’ formula. The output is recorded in Log 5.

Run 6: Model 5 with the second knowledge base with corrected conditional probabilities rules (4.1) is applied to the set of incidents with independent symptoms to provide more knowledge about the use of the complete Bayes’ formula.

Run 7: Model 2 with the second knowledge base with corrected conditional probabilities rules (4.1) is applied to the set of incidents with independent symptoms to provide more knowledge about the use of the complete Bayes’ formula. The output is recorded in Log 7.

B. Runs on the incidents with interdependent symptoms

Run 5: This run uses Model 1 in processing the second set of incidents with dependent symptoms. This application is a violation of the assumption about independence on which Model 1 is built. The output of the run is Log 5.

Run 6: Model 5 with the advanced knowledge base is used in this run. This model assumes incidents with interrelated symptoms. The output, Log 6, compared with Log 5 represents our basis for evaluating the effect of ignoring dependency in the previous models.

4. Realization of the experiments

A knowledge base for experimentation can be acquired in several ways. One way is by the cooperation of a subject matter expert who is willing to help to formulate his knowledge in the formal form required. A second approach is to find an already established knowledge base which satisfies the requirements for the experiments. A third way is to design a machine learning system which can learn from incidents described by an expert or an expert system.

For these experiments, we chose the second approach for establishment of the first knowledge base which we shall refer to as the simple knowledge base. We have used a knowledge base derived from a demonstration base presented in a work by Naylor (1983) and adjusted it to the requirements of our experiments. The knowledge base does not contain any rules for the occurrence of sets of related symptoms given a diagnosis, only rules corresponding to (2.3) and (2.4). The second knowledge base, called the corrected simple knowledge base, is identical to the simple knowledge base, except for the rules (2.4) which have been corrected proportionate in such a way that the conditions:

\[ (4.1) \sum_{i} P(e|i|d|j)=1, \quad \text{for all } j=1..M, \]

are satisfied. In this way any serious bias in the original rules (2.4) may be compensated.

As used in our experiments, these two databases have both a set S with 68 elements, a set D with 91 elements, a set R1 with 520 statistical rules between pairs of symptom and diagnosis and another set R2 of rules between pairs of symptom and NOT diagnosis. Set R2 has also 520 rules. The complete rule set is R=R1,R2).

The third knowledge base, which we call the advanced knowledge base, was obtained by simulation of a learning process working with a set of 3000 incidents with statistically dependent symptoms for which we knew the correct diagnosis for each incident. This set of incidents is also statistically consistent with the first knowledge base.

The learning process used is assumed to work in the following way. When the learner is presented with a new incident, he checks whether it contains an already known complex of symptoms. If the complex is already known, the occurrence of diagnosis/symptom complex pair is added to the number of already known incidents of this type. If the symptom complex has not previously been encountered, the complex is recorded as a new element discovered of E. After the whole set of incidents is inspected, frequencies for all occurring elements of C are available and are the basis for estimating the marginal probabilities P(d|i) and the conditional probabilities P(e|i|d|j) and P(d|i|NOT d|j).

During the learning process, incidents may both expand the learner's knowledge about diagnoses, symptom complexes and probability rules when new combinations are encountered as well as improving and enforcing the quality of existing rules. In addition to the S, D and R sets of the simple knowledge base, the advanced knowledge base contains 2539 probabilistic rules between diagnoses and symptom complexes and a list of 518 different complexes of symptoms.

The knowledge base of Naylor appears in the form of a medical knowledge base. Since we are not testing how good the system is for making medical diagnoses compared with a medical expert, the subject domain of the base has only interest as far as it represents the typical structure of a real domain. The medical quality of the knowledge base has no interest for the investigation reported in this paper. The subject matter has therefore been disguised in order not to distract us from the main purpose of the study. The knowledge base is, however, assumed to be rather typical with no dominating a priori probability. The greatest a priori probability is less than 0.20. The number of symptoms connected to each diagnosis is on average between 5 and 6.

To carry out the experiments, a set of four related systems, GENERATOR, ESTIMATOR, PREDICTOR and ANALYZER, were used. These systems were all developed in PASCAL and are described in detail in other papers [Nordbotten 1988a, 1988b].

GENERATOR was used to generate three sets of incidents. The first set, which contained 3000 incidents with statistically dependent symptoms, was the training set used for the development of the advanced knowledge base by means of ESTIMATOR as described above. The second set gener-
ate had 400 incidents with statistically independent symptoms and finally, the third set which contained another 400 incidents with statistically dependent symptoms. The two last sets were used in the simulation runs.

GENERATOR first built a cumulative probability distribution over the set of diagnoses. From a random number generator, a random number, \( r \), in the interval \( 0 < r < 1 \) was obtained for each case to be generated. This number was used as an argument in the inverse probability distribution which gave one diagnosis as a result. For the second set of statistically independent symptoms the generation was carried out in this way. Each rule which related the generated diagnosis to a symptom was then identified in the knowledge base and by a new random number, the symptom was included in the incident with the probability describing the rule.

The sets of incidents with statistically dependent symptoms were generated by forcing a statistical dependence on the set of symptoms. The rules used in the generation of the interrelated symptoms are given diagnosis \( d[i'] \):

\[
\text{(4.1) } \begin{cases} \text{IF} & P(s[j] | d[i']) = \max. \text{ for all } s[j] \text{ in } S \\
\text{AND} & (a*P(s[j] | d[i']) <= 1) \\
\text{THEN} & 1. \text{ IF } s[j'] = \text{TRUE} \\
& \text{then} \\
& P(s[k] | d[i']) := a*P(s[k] | d[i']) \\
\text{OR} & 2. \text{ IF } s[j'] = \text{FALSE} \\
& \text{then} \\
& P(s[k] | d[i']) := (1-a*P(s[j'] | d[i'])) \times P(s[k] | d[i'])/(1-P(s[j'] | d[i'])) \\
\end{cases}
\]

\[
\text{(4.2) } \begin{cases} \text{IF} & P(s[j'] | d[i']) = \max. \text{ for all } s[j] \text{ in } S \\
\text{AND} & (a*P(s[j'] | d[i']) <= 1) \\
\text{THEN} & 1. \text{ IF } s[j'] = \text{TRUE} \\
& \text{then} \\
& P(s[k] | d[i']) := P(s[k] | d[i'])/P(s[j'] | d[i']) \\
\text{OR} & 2. \text{ IF } s[j'] = \text{FALSE} \\
& \text{then} \\
& P(s[k] | d[i']) := 0. \\
\end{cases}
\]

These rules state that if the most likely symptoms occur, then all the other possible symptoms occur with an increased probability. If the most probable symptom does not occur, the probabilities of the other symptoms are decreased. The dependence is determined by the factor \( a \) and the rules above. The factor is set \( a = 1.33 \) in generation of the set of incidents. The rules maintain the conditional probabilities \( P(s[j] | d[i]) \) for all symptoms which may occur by increasing the probabilities for symptoms other than the most probable when this occurs, and reducing the probabilities when the most probable does not occur.

To check the distribution by diagnosis of the generated test sets, the chi-square values of the frequency distributions for the expected and the generated diagnoses are computed to be 65.4 and 85.3, respectively, with 93 21 degrees of freedom. With a confidence level of 0.95 the critical value is about 115, the computed values will be well within the limit and the generated sets cannot be rejected as drawn from the distribution of the knowledge bases.

ESTIMATOR was used to simulate the learning process forming lists of diagnoses, symptoms and symptom complexes from the set of 3000 incidents generated by GENERATOR. ESTIMATOR also formed the probabilistic rules for the conditional probabilities for symptom complexes given diagnosis which the advanced knowledge base contains.

PREDICTOR is a multiple consultation system shell, which, supplemented with a probabilistic knowledge base, forms a set of five alternative probabilistic consultation systems. It gives the user the option of selecting between the five different inference algorithms corresponding to the inference models discussed above. The system is designed to handle sets of incidents without intervention by the user after initial specification. All results are recorded in log sets permitting study and evaluation of the performance of the respective systems.

ANALYZER is a statistical system especially developed for the processing of log sets from PREDICTOR. The system provides tabulations and computes test statistics for each log set as well as test statistics for pairs of log sets. For each log set, the tabulation includes:

1. Number of incidents
2. Number of correctly predicted diagnoses
3. Number of incorrectly predicted diagnoses
4. Average estimated probability of correct predictions
5. Average estimated probability of incorrect predictions
6. Number of occurring symptoms
7. Number of investigated symptoms
8. Number of detected symptoms
9. Average processing time
10. Efficiency

In addition the chi-square of the differences between frequency distributions for:

11. Predicted and true diagnoses

and the F-statistic for:

12. The variations of the probability estimates for correct and incorrect predictions

are computed.

The average efficiency indicators are computed as:

\[
\text{(4.3) Efficiency}=100 \times \frac{\text{no. of correct predictions}}{500 \times \text{no. of incorrect pred.} + \text{no. of investigated symptoms} + 0.05 \times \text{average processing time}}
\]

The formula states that the penalty of an incorrect prediction is 500 times the costs of investigating a symptom which in turn cost 20 times each time unit of processing time. This measure is rather arbitrary and serves mainly for the purpose of illustration.

The hardware used was a MS-DOS 80286-machine with 512 Kb memory and a 20 Mb hard disk.

DECEMBER 1989 SIMULATION 285
5. Analysis and conclusions

For all the comparisons done, the three test statistics 13, 14 and 15 mentioned above had high significant values, rejecting any hypothesis about equality between the results of the runs carried out. They will not be further commented on.

5.1 Hypothesis 1: Quality and efficiency of two models

We will start by comparing the quality of the results from Model 1 and Model 2 by looking at the number of successful predictions. The two models were analyzed on basis of results from Run 1 and Run 2 on the set of incidents with independent symptoms. The results are summarized in Table 5.1a.

The results indicate that Model 1 has a higher success in predicting correct diagnoses than Model 2. Out of the 400 incidents, 91.8% were correctly predicted by Model 1 compared with only 74.0% by Model 2. The total number of symptoms investigated by Model 1 is almost 60% and by Model 2 about 15% of all possible symptoms.

Table 5.1a: Comparison between the inference Model 1 and 2.

<table>
<thead>
<tr>
<th></th>
<th>Run 1</th>
<th>Run 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of incidents:</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Number of correctly predicted diagnoses:</td>
<td>367</td>
<td>296</td>
</tr>
<tr>
<td>Number of incorrectly predicted diagnoses:</td>
<td>33</td>
<td>104</td>
</tr>
<tr>
<td>Avg. est. probability for correct predictions:</td>
<td>0.956</td>
<td>0.945</td>
</tr>
<tr>
<td>Avg. est. probability for incorr. predictions:</td>
<td>0.888</td>
<td>0.963</td>
</tr>
<tr>
<td>Number of occurring symptoms:</td>
<td>1635</td>
<td>1635</td>
</tr>
<tr>
<td>Number of investigated symptoms:</td>
<td>15955</td>
<td>4174</td>
</tr>
<tr>
<td>Number of detected symptoms:</td>
<td>1576</td>
<td>856</td>
</tr>
<tr>
<td>Average processing time:</td>
<td>52.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Efficiency:</td>
<td>1.095</td>
<td>0.573</td>
</tr>
<tr>
<td>Chi-sq. for diff. of pred. and true diagnosis:</td>
<td>31.2</td>
<td>139.1</td>
</tr>
<tr>
<td>F value (1,398) for prediction probabilities:</td>
<td>0.150</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Model 2 is, however, much less resource-consuming than Model 1. While Model 1 requires investigation of on the average 43 symptoms per successful prediction, Model 2 only investigates 14 symptoms per successful prediction. If we consider the ratio of successful predictions by number of symptoms investigated for the two models, we find that Model 2 has a ratio 3 times higher than that of Model 1.

Actual applications are frequently such that a penalty has to be paid for each incorrect prediction. The penalty can be considered as a cost and added to the number of investigated symptoms. If, for example, the penalty for each incorrect prediction is 187 times the cost of investigating a symptom, Model 1 will break even with Model 2 as to cost. The illustrative efficiency indicator introduced in (4.3) and reported for both models in the table, shows a value for Model 1 which is nearly twice as high as for Model 2. It should be remembered that this efficiency measure is arbitrary and serves here as an illustration.

Another interesting factor in the comparison of the performances of Model 1 and Model 2, is the number of detected symptoms per investigated symptom. Model 1 detected less than 1 symptom per 10 investigated, while Model 2 found more than 2 symptoms per 10 investigated. In this sense the search process of Model 2 is more effective than that of Model 1. But it should be remembered that confirming the non-occurrence of a symptom is as important as detecting an occurring symptom.

In actual situations, the resources required by the inference process itself may be an important component in the cost of carrying out a consultation. One way of measuring the extent of the inference process is to measure the time needed for the inference process. The absolute value of the time will depend on the equipment used, and we will use time units which are comparable for experiments performed within the environment which was available for us.

In the time units used in measuring the processing of these experiments, Run 1 required 52.6 time units on average per incident while Model 2 only used 1.5 time units.

An interesting question is how do the incorrectly predicted diagnoses influence the distribution of incidents by diagnosis. Are they systematically clustered to some diagnoses or are they randomly distributed? The more clustered the incorrect predictions are, the more the distribution of the predictions will deviate from the distribution of true diagnoses. Table 5.1a shows the chi-square value with 46 degrees of freedom for the distribution of predicted diagnoses compared with the distribution of true diagnosis for each run. The values are 31 and 139 for Run 1 and Run 2, respectively. The critical value of the chi-square with 46 degrees of freedom and a probability 0.05 for rejecting a true hypothesis is about 60. If a run produces a chi-square greater than this value, we will reject the hypothesis that the predictions can be drawn from the same distribution as the true diagnoses. Run 1 produces a distribution of predicted diagnoses with a chi-square value within the critical chi-square value, while Run 2 must be rejected as being produced from the same distribution as that representing the true values. Table 5.1b gives a listing of the different frequencies of predicted and true diagnoses in Run 2 for which the difference is >=5.

Table 5.1b: Different frequencies for predicted and true diagnoses in Run 2

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>True</th>
<th>Frequency</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>56</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>31</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>79</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>14</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

As expected the table indicates serious problems for Model 2 in connection with the prediction of particularly the diagnoses 65, 66 and 67. A further study of the reason for this might have given a basis for improving Model 2. This is, however, outside the scope of these experiments.

In summary, Run 1 and Run 2 support the hypothesis that the overall strategy of Model 1 is significantly better than the simpler strategy of of Model 2.
5.2 Hypothesis 2: Impact of the stop rule

To analyze the relative impact of the two stop rules on the results, we use the output from Run 1 and 3 produced by Model 1 and Model 3. Model 3 differs from Model 1 in applying the simpler stop rule (2.9). The results of the runs of the two models are summarized in Table 5.2.

### Table 5.2: Comparison between Run 1 and Run 2.

<table>
<thead>
<tr>
<th></th>
<th>Run 1</th>
<th>Run 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of incidents</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Number of correctly predicted diagnoses</td>
<td>367</td>
<td>261</td>
</tr>
<tr>
<td>Number of incorrectly predicted diagnoses</td>
<td>33</td>
<td>139</td>
</tr>
<tr>
<td>Avg. est. probability for correct predictions</td>
<td>0.956</td>
<td>0.942</td>
</tr>
<tr>
<td>Avg. est. probability for incorr. predictions</td>
<td>0.888</td>
<td>0.965</td>
</tr>
<tr>
<td>Number of occurring symptoms</td>
<td>1635</td>
<td>1635</td>
</tr>
<tr>
<td>Number of investigated symptoms</td>
<td>15955</td>
<td>6447</td>
</tr>
<tr>
<td>Number of detected symptoms</td>
<td>1516</td>
<td>838</td>
</tr>
<tr>
<td>Average processing time</td>
<td>52.6</td>
<td>15.6</td>
</tr>
<tr>
<td>Efficiency</td>
<td>1.095</td>
<td>0.342</td>
</tr>
<tr>
<td>Chi-sq. for diff. of pred. and true diagnosis</td>
<td>31.2</td>
<td>319.1</td>
</tr>
<tr>
<td>F value [1,398] for prediction probabilities</td>
<td>0.150</td>
<td>1.305</td>
</tr>
</tbody>
</table>

Run 1 was discussed in the previous section. The results from Run 3 indicate that the change to the simpler stopping rule has a deteriorating effect on the quality. The number of incorrect predictions is 3 times higher in Run 3 than in Run 1. The number of investigated symptoms per successful prediction is slightly less for Run 3 compared with Run 1 so the investigation cost per successful prediction may be comparable. However, with a penalty for incorrect predictions, the efficiency indicator shows a serious drop in efficiency due to the simpler stop rule which obviously stops investigations much earlier than the more sophisticated stop rule of Model 1. The chi-square value for Run 3 also provides a strong indication that the simple stop rule has a dangerous effect on the distribution of predictions.

The conclusion of the analysis of Run 1 and Run 3 is the hypothesis that the stop rule of Model 1 contributes to a significantly higher quality than the simpler rule of Model 3 and must be maintained.

5.3 Hypothesis 3: Impact of versions of Bayes' formula

The Bayes' formulas (2.1) and (2.2) were used in Models 1, 2 and 3. This formula requires the rule set (2.4) in addition to the first set (2.3). The knowledge embedded in the set (2.4) may not be very precise since it is difficult by informal methods to appraise the conditional probabilities of the different symptoms given NOT the individual diagnoses. The alternative used in Model 4 is to compute the estimates according to the complete formula (2.10) and (2.11). This requires more computation, but on the other hand, does not require the second set of rules (2.4).

The results of Run 4 are summarized in Table 5.3a in which a comparison with Run 2 is made.
recomputed \( P(d[i]) \) and \( (1-P(d[i])) \).

The results reviewed in this section indicate the hypothesis that the use of the complete Bayes' formula contributes to high quality and efficiency. However, if the probability rules satisfy the conditions (4.1), this may not be true.

5.4 Hypothesis 4: Impact of interdependent symptoms

The most serious objection to many of the consultation systems using Bayes' Theorem is the violation of the assumption of statistical independence between the symptoms. To avoid this violation, we have in our experiments so far used a test set of incidents which was constructed in such a way that the assumption about independence was satisfied. Such a set could be used for the explorations so far since it put the systems through a test which was at least as difficult as one which is based on a set of incidents with interdependent symptoms.

To investigate the implications of violating the assumption of independence between the symptoms in an incident, we used the second set of incidents which was generated in such a way that a dependence between symptoms was introduced, and the third advanced knowledge base which also contained rules with probabilities for the occurrences of complexes of symptoms. Run 5 was carried out by Model 1 and is an intended violation of the assumption about independence on which this model was constructed. Run 6 is by Model 6 which is constructed on the assumption that the symptom complexes are statistically independent.

Comparison of the results from the two runs should therefore indicate the effects of the violated assumption of statistical independence between the symptoms.

The results are summarized in Table 5.4. We note that the number of incidents is less than 400. The explanation is that in generating 400 incidents 24 of them came out with no symptoms. We decided to eliminate them from our study. The rates of successful predictions are therefore 92.6% and 78.2% for Run 1 and Run 5 respectively. The reason for the difference is that Model 1 and 5 share the search process, the stop rule, but Model 5 recalculates the probabilities for the diagnosis according to the advanced knowledge base and then selects the most probable prediction based on these probabilities. Because of the limited number of incidents used for developing the advanced knowledge base, some rules are based on few observations and the estimated conditional probabilities can be uncertain. The detected complex may therefore lead to a different prediction than that of Model 1. The above differences should be expected and are of less interest in this connection.

The interesting factors in this comparison are the estimated probabilities of the correct and the incorrect predictions and the F value. These show a significant difference between Run 5 and 6 to which we shall return in the subsequent section.

We conclude that Runs 5 and 6 do not support the hypothesis that there is a significant difference in the performance of Model 1 and Model 6 on a set of incidents with statistically dependent symptoms. There are, however, other differences to which we return in the next section.

5.5 Hypothesis 5: Estimated probabilities as quality measures

In actual applications, true values for the conclusions will of course not be available. Are the estimated probabilities of the predictions useful indicators of the reliability of the predictions?

In two runs, Run 2 and Run 3, the average probabilities for the incorrect predictions are higher than the average probabilities for the correct predictions. The difference between the average probabilities for correct and incorrect predictions is less than 10% for Run 1 and Run 5, both based on Model 1. We can conclude that for Model 1, Model 2 and Model 3, the estimated probability has no value in evaluating the reliability of a prediction.

We have computed the F-statistic to test the hypothesis that there is no difference between the probability estimates for the correct and the incorrect predictions. The critical F value for a confidence coefficient 0.95 and \( 1, 400/376 \) degrees of freedom, is about 3.8.

For Model 4 we can conclude that the F-test gives no basis for rejecting the hypothesis of no difference between the average probabilities for correct and incorrect predictions. The F value is, however, great enough to use the estimated probability as a guide. Model 5 has a F-value which indicates significant difference between the correct and the
incorrect predictions, and its estimated probabilities can
without significant risk be used as reliability indicators.

The conclusions just drawn supplement the discussion of
the previous section. The theoretically more justified Model
is the only one which provides probability estimates
which have some value. The reason seems to be that all the
other models are based on the assumption of statistically in-
dependent symptoms which, in actual applications, likely to
be violated.

5.6 Summary of conclusions and experimental experience
We can summarize our answers to the problems we
wanted to investigate in this way:

1) In actual applications, the more sophisticated Model 1
can be expected to give more successful predictions than
Model 2.

2) The stop rule for selecting a prediction when the
estimate of the probability of the prediction has reached
a certain predetermined level, even if this level is 0.99, is
not satisfactory.

3) The complete Bayes’ formula contributes to more
successful predictions than the use of the simpler
formula.

4) Prediction probabilities are overestimated in models
which do not assume intersymptom dependence.

5) In models which ignore the intersymptom dependence,
the probabilities of the predictions cannot be expected to
be useful indicators of reliability.

During the execution of the experiment, the high compu-
tational workload was felt and some computational sessions
required hours of uninterrupted efforts by the hardware.
Faster equipment would obviously be needed if more
extensive experiments should be carried out.

5.7 Some further work
The results discussed above indicate that more should be
done to explore the characteristics of probabilistic consulta-
tion systems. A major problem in current models is to
develop an indicator of prediction reliability. Other possible
inference models should be developed and studied. This
will require further development of systems for automatic
knowledge acquisition.

The ignorance of dependencies between symptoms
related to a diagnosis is a problem which deserves further
study. So far few, if any, knowledge bases which take the
dependence into account have been constructed because of
the difficulty in acquiring the necessary detailed and
complex knowledge. However, with certain computerized
learning models, it is possible to acquire the necessary
knowledge if we can get access to training material of
recorded incidents treated by experts in the respective
domain.

Even if the needed knowledge is available, it is a question
whether this additional knowledge may be so uncertain that
a more sophisticated inference might not represent any im-
provement in practical work.

One aspect of the computerized statistical inference
models is the time spent in resolving each task. Statistical

6. Acknowledgement
This paper has been prepared during a stay as a visiting
faculty member with the University of Hawaii, Department
of Information and Computer Sciences, which provided a
professional environment. The visit was made possible by
the financial support from the University of Bergen, the
Royal Norwegian Council for Research, and the Norwegian
Research Council for Social Sciences.

The following colleagues made constructive comments to
the paper: Associate Professor Joan C. Nordbotten, who, as
usual, is my chief critic. Dr. Jiri Ivanek, who stayed as a
visiting colleague with my department during the spring
term of 1987 and made some constructive comments to my
ideas on inference models at an early stage. The author is,
however, solely responsible for points not taken and for any
errors in the paper.

I am very grateful to all these organizations and persons
that made it possible for me to carry out the work reported
in this paper.

7. References
Naylor, C.. Build Your Own Expert System, Halsted Press, a Division of John
Wiley and Sons, NY, 1983.
Naylor, C.. "How to Build an Inference Engine", in Forsyth,R. (ed.): Expert
Nordbotten, S.. "Experimenter—An Experimental Environment for
Knowledge Base System Simulation", (in press).

Svein Nordbotten is currently Dean of
the Social Science Facility at the Univer-
sity of Bergen, Norway as well as Vice
President of the International Statistical
Institute. Previously, Nordbotten was the
Associate Director of the Central Bureau
of Statistics of Norway and Director of
the United Nations Statistical Office in
New York. From 1983-1985 he served as
President of the International Association
of Statistical Computing and a statistics
and computer consultant.