THE FORMAL DESIGN AND EVALUATION OF A VARIETY OF MEDICAL DIAGNOSTIC PROGRAMS

by

B.S. Todd and R. Stamper

Technical Monograph PRG-109
ISBN 0-902928-86-4

September 1993

Oxford University Computing Laboratory
Programming Research Group
11 Keble Road
Oxford OX1 3QD
England

Oxford University Computing Laboratory
Programming Research Group
11 Keble Road
Oxford OX1 3QD
England

Abstract

One of the most successful and most widely studied applications of computer-aided diagnosis is that of acute abdominal pain. However, widespread introduction of computers into clinical practice seems to be hindered by their limited diagnostic accuracy. This monograph documents some experiments we have carried out to investigate the effect on diagnostic accuracy of using various statistical and knowledge-based methods to take dependencies between clinical observations into account. We present detailed specifications of a variety of statistical and knowledge-based programs within a common formal framework using the Z specification language. We describe how we collected a retrospective database of 1270 cases of abdominal pain of suspected gynaecological origin. This we used to evaluate all methods.

Our results show that no significant improvement in accuracy can be made by taking interactions into account; independence Bayes is optimal in this application. However, the nearest neighbours method using a new metric appears to be at least as accurate. The metric is the Euclidean distance between the posterior probability distributions over the possible diseases, computed using the independence Bayes formula. We argue that the nearest neighbours method is more suitable for clinical use than the direct application of independence Bayes because of improved accountability. The computer analysis is encoded and displayed to the user as a small set of *actual* cases that presented similarly to the new case. The user can retrieve these cases and inspect both their presentations and outcomes if he wishes. The ability to justify and support decisions in this way should be invaluable in safety-critical fields such as medical diagnosis.

Contents

Chapter 1

Introduction

In this chapter we give an overview of the experiments we have carried out to investigate the effect an diagnostic accuracy of using various statistical and knowledge-based methods to take dependencies between clinical observations into account. We start by briefly surveying the field of computer-aided medical diagnosis, with the emphasis on the use of independence Bayes in the diagnosis of acute abdominal pain since this is the most widely studied and most successful application. Widespread introduction of computers into clinical practice seems to be hindered by their limited diagnostic accuracy. Can accuracy be improved by taking interactions into account? We address this question by designing and implementing a variety of statistical and knowledge-based programs, and testing them on the same data set of patients. Our results suggest that independence Bayes is optimal, but that the nearest neighbours method using a new metric is at least as accurate. We argue that the new nearest neighbours technique is more suitable than independence Bayes for clinical use. However, it seems unlikely that accuracy can be improved by new computational methods.

1.1 Computers and Medical Diagnosis

Ever since the electronic computer became commercially available, clinicians have been interested in its potential to assist medical diagnosis [Led59, Lip61]. The motivation for machine assistance is that doctors are not fully aware of the significance of the observations they make; numerous studies [Dom78, Kni85, Lav90] have since confirmed this. Perhaps the best known and most successful application of a diagnostic program is due to de Dombal and colleagues in Leeds, England [Dom72]. Bayes theorem was used to classify cases of acute abdominal pain into one of seven mutually exclusive disease categories. Since the method involves an assumption that observations (symptoms and signs) are conditionally independent within each disease category, the method is often referred to as 'independence Bayes'. The computer analysis is regarded as a supplemental test much like any other investigation [Dom84].

Over the last twenty years, independence Bayes has been used to assist the diagnosis of acute abdominal pain in an enormous number of patients worldwide [Dom91]; a multicentre trial in the UK alone involved 16737 patients [Ada86]. The accuracy of the computer program, however, was now less than the initial studies had suggested. Furthermore, although

clinicians made fewer errors when the computer system was introdnced, it became clear that much of the improvement was due to the discipline of using structured data-collection forms to record patient histories [Wel89, Gun91]. A re-analysis of 5193 of the cases showed that most of the observed improvement was actually due to 'non-specific abdominal pain' (NSAP) being diagnosed rather than no diagnosis being made [Wel92]. Disturbingly, it was noted that if the computer made any diagnosis other than NSAP, there was less than a 50% chance of it being right!

A Scandinavian study [Fen87] reported a similar effect of introducing structured data-collection forms alone: the negative laparotomy rate decreased from 26% to 16%, and the positive laparotomy rate for appendicitis increased from 69% to 77%. However, computer accuracy was less in this study than in the UK multicentre trial. Other groups have also experienced difficulty in achieving a computer accuracy equal to that of the clinician (e.g. [Kir87, Sut89b, Fla92]. Indeed, in one retrospective study of 200 cases of acute abdominal pain seen in an accident and emergency department, initial clinical accuracy was recorded as high as 65% [Mai88]. This is similar to other workers' results with computers and structured forms. In another study of 158 cases of acute abdominal pain admitted under one surgical firm, use of structured forms and selective use of diagnostic laparoscopy led to only 3 management errors (two unnecessary appendicectomies and a laparotomy for diverticular abscess); when independence Bayes was applied retrospectively to the same cases, adoption of the computer diagnosis would have produced a total of 26 similar errors [Pat89]. The value of the computer analysis is therefore unclear, especially as there are many other diagnostic techniques available and perhaps underused (e.g. ultrasonography and fine catheter aspiration) [Pat91]. In the West Lothian study of computer-aided diagnosis of acute abdominal pain, computer accuracy gradually fell from 78.5% to 55.0% over a 15 year period, while clinical accuracy in the accident and emergency department remained fairly constant at 60.4% to 70.0%. A decision was therefore made to withdraw computer-aided diagnosis and continue with only structured data-collection forms [Sto92]. Similar experiences have been reported in other areas of medicine: Engle *et al* concluded that after 30 years of research into computer-aided diagnosis of haematological diagnosis they have finally accepted that the 'intelligent' computer does not even seem useful as an aid [Eng92]. Low computer accuracy seems to limit widespread use of computers as decision aids in medicine [Sut89a]. Therefore, can accuracy be improved by using computational methods other than independence Bayes?

1.2 Statistical Interactions

Any possibility of improvement depends crucially on the presence of statistical interactions between observations within each disease category [Nor75a]; if no such interactions occur then clearly the independence model is optimal. Experience, however, suggests that some symptoms and signs do tend to eclipse others (e.g. [Dix91]). Indeed, numerous studies have shown the presence of interactions in a variety of applications [Nor75b, Fry78, Tod93a]. However, it does not necessarily follow that taking these interactions into account will improve accuracy. Hilden [Hil84] has pointed out that for a particular class of probability distributions, a classifier based on independence Bayes is optimally accurate despite statistical interactions. Furthermore, for some of these distributions, the independence Bayes formula even computes the correct posterior probabilities! However, in practice indepen-

dence Bayes is usually found to poorly calibrated, producing over-optimistic estimates of the posterior probabilities (e.g. [Rus83, Gup89, Wel92]). Poor calibration is easily rectified [Dom92]. The central question which this monograph addresses is whether statistical models other than independence Bayes are more accurate classifiers.

Some early studies [Nor75a, Fry78] suggested that accuracy can be improved by taking interactions into account. However, in both cases the test set had been used for training purposes; a classifier with more degrees of freedom would naturally be better able to separate examples from different classes. Other comparative studies in the diagnosis of thyroid disease [Nor71], diagnosis of liver disease [Cro74], prediction of outcome of head injuries [Tit81], and prognosis of heart disease [Rus83] have generally concluded that independence Bayes is the most suitable classifier. In a study of patients with abdominal pain presenting to general practitioners, a six-page diagnostic flowchart, a linear discriminant and independence Bayes were found all to be of similar accuracy, and much worse than the clinicians [Ori86]. In a more recent study concerning 6,000 patients with acute abdominal pain, independence Bayes was found to be significantly more accurate than another method (induced decision tree) which avoids the independence assumption [Gam91]. Consequently, many regard independence Bayes as optimal [Edw84], and some regard discussion about the type of computer program to use as outmoded [Dom91].

However, a French study involving 6916 patients with acute abdominal pain reported in increase in computer accuracy from 63.7% to 67.7% as a result of taking into account pairwise interactions by means of a Lancaster model [Ser86]. Some other studies have also reported improvements. Neural networks were found to be more accurate than independence Bayes for discriminating appendicitis from NSAP [Ehe91]. The test set was small though (3 cases of appendicitis and 61 of NSAP), and the difference in accuracy does not appear to be statistically significant. Similarly, Emparanza et al [Emp88] found logistic discrimination to be more accurate (91%) than independence Bayes (84%) for the same discrimination task in children (training set 569, test set 100). However, ROC curves were not shown, so it is conceivable that the observed improvement was due to the better calibration of the logistic discriminant. In another application, the prediction of recurrent upper gastrointestinal bleeding, (again a binary discrimination task) an increase in accuracy from 57% to 67% was achieved by taking interactions into account by means of a dependence tree [Ohm88]. However, interpretation is difficult because the training and test samples were small (322 and 207, respectively) and were collected over different study periods between which management policy changed. In summary, therefore, while it is clearly evident that statistical interactions are usually present amongst observations, it is less evident that accuracy can be improved by taking these interactions into account.

Perhaps the reason is that statistical methods more sophisticated than independence Bayes tend to require estimation of many more numerical parameters. The consequent tendency to overfit to the training sample offsets any gain in accuracy that can be achieved by taking interactions into account. Perhaps current databases are simply too small for purely statistical methods to exploit interactions fully. One possible solution is to exploit available knowledge of causal mechanisms in the particular domain of application in order to construct a classifier that is less highly parameterized. The knowledge can be expressed as inference rules (e.g. [Dav71]), or as an explicit causal model (e.g. [Hai88, Hec92b]).

Rule-based systems have now been compared with independence Bayes in several applications. One of the earliest was the diagnosis of dyspepsia: no significant difference in accu-

racy was found [Fox80], although the inference rules do not appear to have been equipped with numerical certainty factors. The rule-based approach was found to be less accurate than independence Bayes in the diagnosis of lymph node pathology [Hec92a], and about as accurate as independence Bayes in the diagnosis of acute abdominal pain [Dom89]. However, if subjective certainty factors are used then any gain in accuracy achieved by taking interactions into account is offset by errors in estimation of the parameters; numerical estimates elicited subjectively from clinicians are well-known to be frequently inaccurate [Lea72, Kni85, Cha87]. Few rule-based expert systems appear to estimate numerical parameters objectively from an actual training sample. A recent exception is EMERGE [Hud91] currently being developed for the diagnosis of chest pain.

An alternative to the rule-based approach is the construction of an explicit statistical model of the causal mechanism by which diseases produce their manifestations. The most general and best-known representation is that of the Bayesian network (e.g. [And91, Hec92b]). Cooper asserts that if causal and probabilistic knowledge is available then the causal graph (Bayesian network) method will generally yield more accurate diagnostic results than independence Bayes [Coo86]. Ludwig and Heilbronn, however, reported that a probabilistic causal graph was less accurate than simple logistic regression for the diagnosis of chest pain [Lnd83]. Nevertheless, intuition supports Cooper's view that a carefully constructed knowledge-based program should be more accurate than independence Bayes, provided that

1. all numerical parameters are estimated objectively from a random training sample, and
2. proper attention is paid to constraining the number of degrees of freedom so that overfitting does not occur.

The rest of this monograph documents a series of experiments we have carried out to test this hypothesis.

1.3 Methods

We designed and implemented a set of diagnostic programs embodying various paradigms, ranging from the purely statistical to the knowledge-based. The latter included categorical flowcharts, rule-based systems and Bayesian networks. Some of the methods are standard, while others are innovative. However, we describe *all* methods within the same formal mathematical framework using the notational conventions of the Z specification [Spi88]. We do this principally for two reasons: to document precisely the methods that we have implemented and to reveal relationships between the various methods. However, the process of formal specification can also be a source of unexpected insight. For example, the concept of an iterative flowchart (Chapter 7) arose naturally as a way of producing idempotent inference procedures from individual inference steps.

Recently software engineering techniques have been increasingly applied to knowledge-based systems. Formal specification has been used to present the theoretical basis for models of reasoning, such as rule-based deduction [Bez91] and Bayesian classification [Pen89]. It has also been used to provide precise descriptions of particular inference systems [Hai88]. However, the style of specification varies widely. This is because the specifications are of existing systems, whose assumptions partially determine the style of specification. Yet one

of the strengths of mathematical specification is its ability to abstract away from details of particular implementations. This is achieved by starting with an abstract view of inference that is so general that it should apply to almost all systems. This model is then refined by making explicit design decisions until a particular implementation is reached. This is the style we have adopted in this monograph. We use simple mathematical definitions (sets, sequences, functions etc.), interleaved with prose which motivates and justifies the design decisions.

In order to evaluate the programs we compared them on the same data set so that pairwise tests of statistical significance of any observed difference could be applied. The application we chose was the diagnosis of abdominal pain because it has been so widely studied, and it appears to be the most successful yet. In particular, we confined our attention to a specific subgroup of patients, those for whom the pain is suspected to be of gynaecological origin. The scope for exploiting knowledge of causal mechanisms appears to be greatest in this selected group of patients. We collected our own data because we needed as detailed information as possible about the pathophysiological state of each patient in order to estimate the various numerical parameters for the knowledge-based programs. Since we were interested in comparing the different programs themselves rather than in comparing computer with clinician, retrospectively collected cases were adequate for our purpose since any impairment of the quality of clinical data tends to disadvantage all programs similarly. We collected a total of 1270 cases, the first 202 of which could not be used for testing any of the knowledge-based programs because those cases had been used to assist knowledge-base construction. We tested the programs on all the remaining 1068 cases, using a cross-validation strategy to avoid bias when estimating all numerical parameters.

The evaluation measure we used universally is the overall crude error rate, taking the disease with highest calculated posterior probability as the computer's diagnosis. This enabled us to compare all programs, including a flowchart whose output is categorical, on a common scale. We did not explore other measures such as the quadratic or logarithmic score [Tit81] because we felt this was too sensitive to the calibration of the program, and would be likely to unfairly disadvantage independence Bayes in particular. Poor calibration can always be improved by quite simple means (e.g. a look-up table), so measures such as the quadratic score are potentially misleading. Nor did we attempt to weight errors with associated costs: since we included 19 diseases in our application, this would have involved subjective estimation of 342 utilities! However, we do present full discrimination matrices for all our programs so that sensitivity, specificity and reliability can be calculated with respect to each condition.

Our results show that no significant improvement in accuracy can be made by taking interactions into account; independence Bayes is optimal in this application. However, the nearest neighbours method using a new metric appears to be at least as accurate. The metric is the Euclidean distance between the posterior probability distributions over the possible diseases, computed using the independence Bayes formula. We argue that the nearest neighbours method is more suitable for clinical use than the direct application of independence Bayes because of improved accountability. The computer analysis is encoded and displayed to the user as a small set of *actual* cases that presented similarly to the new case. The user can retrieve these cases and inspect both their presentations and outcomes if he wishes. The ability to justify and support decisions in this way should be invaluable in safety-critical fields such as medical diagnosis.

Chapter 2

A Formal Model of Diagnostic Programs

This chapter introduces a simple abstract model of inference which will serve as a common starting point for the design of all diagnostic programs described in this document. As an illustration, the model is refined to a simple non-parametric statistical program.

2.1 A Diagnostic Program

A *diagnostic program* assists the interpretation of clinical findings. It takes as input any information that has been gathered regarding a patient, and outputs all conclusions that can be drawn and all inferences that can be made. We therefore regard a diagnostic program (DP) as a function that manipulates information of clinical relevance.

$$DP \triangleq Info \nrightarrow Info \tag{2.1}$$

The program's input generally consists of symptoms and other historical items, physical and radiological signs, the results of certain other investigations, and perhaps details too of any established medical condition the patient is known to have. The program's output is a more complete description of the patient: one which includes a diagnostic assessment of the clinical findings.

Notice that a diagnostic program is not necessarily a total function. If a patient description is illogical, or impossible for any reason, then the result of applying a diagnostic program to such a description is undefined. (For simplicity, we regard the purpose of a diagnostic program to be the diagnosis of disease in patients rather than the related task of detecting errors in patient descriptions.) Therefore, the domain of a diagnostic program consists of precisely the patient descriptions that are feasible; those that are impossible are excluded.

2.2 Information

The diagnostic task amounts to reconstructing a complete description of a patient's medical state from only partial information. The complete medical state of a patient includes all

diseases that are present, all symptoms and signs that are exhibited, all other measurable physiological parameters, and all historical items of clinical relevance including personal details such as sex and age. We refer to a complete description of a patient's medical state as a *case*, and we denote the set of all such descriptions (whether feasible or infeasible) by the symbol CASE.

Medical observation and medical reasoning, however, are pervaded by uncertainty. Rarely is it possible to say without any shadow of doubt what the correct diagnosis is, and sometimes it is not even clear what symptoms and signs the patient has [Gil73]. Therefore, let us represent *information* about a patient as a probability distribution over cases.

$$\text{Info} \mathrel{\widehat=} \mathbb{D}\,\text{CASE} \tag{2.2}$$

A probability distribution over any countable set T is defined by

$$\mathbb{D}T \mathrel{\widehat=} \left\{ d : T \to \text{Pr} \mid \sum_{t:T} d(t) = 1 \right\} \tag{2.3}$$

where 'Pr' denotes the closed interval between 0 and 1.

$$\text{Pr} \mathrel{\widehat=} \{ r : \mathbf{R} \mid 0 \leq r \leq 1 \} \tag{2.4}$$

In practice, there will be only a finite number of possible patient descriptions, since it is not useful to record arbitrarily small variations between cases. We therefore assume that CASE is both finite and non-empty.

The complete absence of any information is represented by the uniform distribution. This is because if there is no reason to prefer one case as more typical than another then the *Principle of Indifference* dictates that every case is assigned equal probability [Nea89]. We denote vacuous information by the symbol $\emptyset$.

$$\emptyset \mathrel{\widehat=} \lambda c : \text{CASE} \bullet \frac{1}{\#\text{CASE}} \tag{2.5}$$

Thus $\emptyset$ is the information we have about a particular patient if we have no knowledge of the patient's medical state, no knowledge of the population from which the patient was drawn, and no knowledge even of the medical properties of different populations: if in short we know nothing either about the patient or about medicine.

2.3 An Abstraction Function

If I denotes the information that a given patient has been drawn randomly from a particular population (i.e. I is the prior distribution) then let us use I to construct a diagnostic program $\mathcal{D}(I)$. Let J be the information obtained about the patient by the clinician. The combined information, taking into account both I and J, is given by the Bayes product of

the two distributions.

$$\mathcal{D} : \text{Info} \rightarrow \text{DP} \tag{2.6}$$

$$\begin{array}{l} \forall I : \text{Info} \bullet \\ \quad \mathcal{D}(I) = \lambda J : \text{Info} \mid I \otimes J \neq 0 \bullet \\ \qquad\qquad \lambda c : \textsc{Case} \bullet \dfrac{I(c) \times J(c)}{I \otimes J} \\ \underline{\textit{where}} \\ \quad I \otimes J \;\widehat{=}\; \displaystyle\sum_{c:\textsc{Case}} I(c) \times J(c) \end{array}$$

Information J is feasible (and the Bayes product is defined) precisely when it is consistent with the complete description of at least one case that can be found in the given population. That is to say, there exists a case c such that $I(c) \neq 0$ and $J(c) \neq 0$.

2.3.1 Some Properties of Vacuous Information

Several simple and intuitive results follow immediately from the definitions above; they are stated without proof. Firstly, no matter what diagnostic program we build, we can always apply it to the vacuous body of information ($\emptyset$).

Lemma 1

$$I : \text{Info} \vdash \emptyset \in \text{dom}\,\mathcal{D}(I)$$

Furthermore, if we take advantage of this facility, since we snpply no information at all about the patieut we wish to diagnose, all we obtain is a description of the random case.

Lemma 2

$$I : \text{Info} \vdash \mathcal{D}(I)(\emptyset) = I$$

Lastly, the diagnostic program constructed from vacuous information is simply the identity function: if we provide our diagnostic program with no information about the general population, it can never infer anything.

Lemma 3

$$\vdash \mathcal{D}(\emptyset) = \text{id}$$

2.4 A Simple Diagnostic Program

In order to estimate the prior distribntion of cases, we sample the relevant population by collecting a sequence of training cases C.

$$C : \text{seq}\,\textsc{Case} \tag{2.7}$$

Provided that the sample C is sufficiently large, the prior distribution is approximated by the relative frequency with which any given case occurs in the sample. Better estimates are obtained by smoothing; for this we use the formula suggested by Cestnik [Ces90] where smoothing coefficient m is strictly positive. The formula introduces small constants into the numerator and denominator based on the default assumption that in the absence of evidence to the contrary (e.g. C is empty), probability is distributed uniformly amongst all cases. Let I_S be our estimate of the prior distribution. (The subscript S denotes the simplicity of the method, and distinguishes it from the more complicated approaches in subsequent chapters.)

$$I_S \;\widehat{=}\; \lambda c : \textsc{Case} \bullet \frac{\#C \upharpoonright \{c\} + \dfrac{m}{\#\textsc{Case}}}{\#C + m} \tag{2.8}$$

The smoothing embodied in Equation 2.8 ensures that the corresponding diagnostic program $\mathcal{D}(I_S)$ is a total function: no input is rejected as infeasible.

Lemma 4

$$\vdash \mathrm{dom}\,\mathcal{D}(I_S) = \mathrm{Info}$$

Notice also an empty training sample conveys no information at all.

Lemma 5

$$C = \langle\rangle \vdash I_S = \emptyset$$

Therefore the resulting program is simply the identity function (Lemma 3); nothing new can be inferred if there are no training examples.

2.5 Specifications

Although in principle a clinician may be prepared to quantify uncertainty in his observations in terms of probabilities, in practice observations tend to be categorical statements. Categorical information about a patient can be regarded as a specification that the patient meets. A specification (Spec) is conveniently identified with the set of all cases that meet the specification.

$$\mathrm{Spec} \;\widehat{=}\; \mathbf{P}\,\textsc{Case} \tag{2.9}$$

A specification s is stronger than another t precisely when s is a subset of t. The strongest specification of all is the empty set: no patient meets this. The weakest specification is the set of all cases 'CASE': all patients meet this. Accordingly, we will refer to the empty set ($\{\}$) as the *impossible* (unsatisfiable) specification, and to CASE as the *universal* specification.

The assertion that a given patient meet a satisfiable (non-empty) specification s, associates zero probability with every case that does not meet s, and distributes probability uniformly

amongst cases which do meet s (Principle of Indifference). Let $\mathcal{K}(s)$ denote the information conveyed by the assertion that s is met.

$$\mathcal{K} \;\widehat{=}\; \lambda s : \mathrm{Spec} \mid s \neq \{\} \bullet \lambda c : \textsc{Case} \bullet \begin{cases} \frac{1}{\#s} & c \in s \\ 0 & c \notin s \end{cases} \tag{2.10}$$

Notice that the assertion simply that a patient meets the universal specification, as one would expect, conveys no information whatsoever because everyone meets that specification.

Lemma 6

$$\vdash \mathcal{K}(\textsc{Case}) = \emptyset$$

2.5.1 A Difficulty

Unfortunately, program $\mathcal{D}(I_S)$ is rather *too* naive. It can assist in the diagnosis of a new patient only if there is some previous case in C which meets the new patient's specification. Otherwise, nothing new can be inferred.

Lemma 7

$$s : \mathrm{Spec} \mid s \neq \{\} \wedge (s \cap \mathrm{ran}\, C) = \{\} \vdash \mathcal{D}(I_S)(\mathcal{K}(s)) = \mathcal{K}(s)$$

This severely limits the applicability of the program. If patient descriptions involve more than a few symptoms and signs, the training sample would need to be astronomically large before such a program were of any practical use.

Clearly, suitability of this method depends on the type of application. One system that implements a similar principle to the one above is TOD ('Time-Oriented Database') [Wey75, Fri86]. This assists the management of rheumatology patients by making prognostic forecasts. At the time of the original report, it contained details of 5500 consultations. Only very weak specifications, however, can be entered. For example, asserting that the patient is female and has systemic lupus erythematosus with proteinuria and increasing ESR (erythrocyte sedimentation rate) led to the retrieval of just 36 training cases from which prognostic inferences could be made.

Chapter 3

Statistical Models

This chapter introduces the idea of using knowledge about the structure of the prior distribution in order to estimate prior probabilities more reliably. The knowledge representations considered are conventional Bayesian networks and exemplar models.

3.1 Background Knowledge

The method for constructing diagnostic program $\mathcal{D}(I_S)$ that was described in the previous chapter makes inefficient use of training data, and consequently it tends to require infeasibly large training samples. More efficient use of training data can be made if we have some knowledge about the structure of the prior distribution (or if we are prepared to make some assumptions). This avoids estimating as many statistical parameters as in the full multinomial case. Furthermore, we can relax the constraint that all details about the medical state of each case in the training sample must be recorded: this is an unrealistic requirement because in practice only partial information is generally available about any given patient.

3.1.1 Variables and Values

In order to discuss the properties of the prior distribution we need to know more about the nature of case histories. In the context of medical reasoning, we are concerned with variables such as 'age', 'site of pain' and 'diagnosis', and the possible values they may take (for example, '24 years old', 'central abdomen', and 'appendicitis'). Not all values are meaningful for any given variable: for example, it would make no sense to talk of the 'site of pain' as being '24 years old'. Let Θ be a relation between variables and their permissible values.

$$\Theta : \mathrm{Var} \leftrightarrow \mathrm{Val} \tag{3.1}$$

By definition, a variable must have some meaningful values. Therefore every variable is in the domain of Θ.

$$\mathrm{dom}\,\Theta = \mathrm{Var} \tag{3.2}$$

The medical state of a patient is described by specifying the values taken by the variables. The case is feasible only if all values assigned are legal.

$$\textsc{Case} \;\widehat{=}\; \{c : \mathrm{Var} \rightarrow \mathrm{Val} \mid c \subseteq \Theta\} \tag{3.3}$$

A possibly incomplete account of a patient's medical state is defined similarly as a partial function. It is denoted by lower case letters.

$$\mathrm{Case} \;\widehat{=}\; \{c : \mathrm{Var} \nrightarrow \mathrm{Val} \mid c \subseteq \Theta\} \tag{3.4}$$

Example 1 *For example, suppose the entire vocabulary of variables and values were limited to just the following.*

$$\begin{aligned} \mathrm{Var} &= \{sex, cough, disease\} \\ \mathrm{Val} &= \{male, female, productive, appendicitis, ureteric\_colic, none\} \end{aligned}$$

A suitable range of permissible values would be

$$\begin{aligned} \Theta \;\widehat{=}\; \{&sex \mapsto male, \quad sex \mapsto female, \\ &cough \mapsto productive, \quad cough \mapsto none \\ &disease \mapsto appendicitis, \quad disease \mapsto ureteric\_colic, \quad disease \mapsto none\} \end{aligned}$$

The case (C1) which is male and has both a productive cough and appendicitis is

$$C1 \;\widehat{=}\; \{sex \mapsto male, \quad cough \mapsto productive, \quad disease \mapsto appendicitis\}$$

Similarly, the case (C2) which is female and has no cough is

$$C2 \;\widehat{=}\; \{sex \mapsto female, \quad cough \mapsto none\}$$

Lastly, the vacuous case (C3) about whom nothing has been recorded is

$$C3 \;\widehat{=}\; \{\}$$

3.1.2 A Bayesian Network

A *Bayesian Network* [Nea89] represents a joint distribution by decomposing it into a chain (sequence) of conditional probability tables, one for each variable. A variable is said to be *anterior* to another precisely when it appears earlier than the other one in the sequence. Each table specifies the conditional distribution of the corresponding variable given all possible states of the anterior variables. With respect to any variable v, the *parents* of v are a minimal subset of anterior variables which exhaust all evidence provided by the anterior variables about the state of v. Only the parents of a variable need to be included in the variable's conditional probability table. Typically this leads to a very large reduction in the size of the resulting table. Greatest savings are made if the chain decomposition corresponds to the direction of physical causation. A variable's parents then consist only of those representing its direct physical causes. Choice of chain decomposition, and selection of parents thus provides a means of representing background domain knowledge, and at the same time allowing more efficient use of training data.

Let P relate each variable to its parents. For any variable v, $P(\!|v|\!)$ is the set of parents of v. Since every parent of v lies strictly anterior to v in some fixed total ordering, P is an *irreflexive acyclic relation*.

$$P : \mathrm{Var} \leftrightarrow \mathrm{Var} \tag{3.5}$$

$$P^+ \cap \mathrm{id} = \{\}$$

The joint distribution (information) specified by a Bayesian network is defined by the product of the conditional probability tables [Nea89].

$$I_B = \lambda c : \textsc{Case} \bullet \prod_{v:\mathrm{Var}} p(v = c(v) \mid P(\!|v|\!) \lhd c) \tag{3.6}$$

For any variable v, value u, and partial case c specifying a state of v's parents, the term $p(v = u \mid c)$ is the conditional probability of v having value u given c. We estimate this from our training sample. However, since Equation 3.6 entails only probabilities conditioned on the state of variables' parents, we can relax the requirement that every training case is complete. In a practical application it is unlikely that all variables will be recorded in every training example. Therefore, formally we override the earlier declaration of training sample C (Equation 2.7) with a new one which allows cases to be incompletely recorded.

$$C : \mathrm{seq}\,\mathrm{Case} \tag{3.7}$$

We take as our estimate of $p(v = u \mid c)$ the relative frequency with which v takes value u amongst the training cases whose observations match those of c. However, we require the sum of the probability estimates to equal unity for all u. Therefore we confine our attention to training cases in which v actually has some recorded value. If no such training cases are to be found, then it is necessary to make a suitable default assumption about the value of $p(v = u \mid c)$. We follow Cestnik's suggestion [Ces90] and take the prior probability $p(v = u \mid \{\})$ as the default. If c is already empty, then the Principle of Indifference dictates that the uniform distribution over the possible values of v is the default. A smooth transition to the default as sample size diminishes is achieved by including small quantities in both the numerator and the denominator of the estimate. We apply Cestnik's formula [Ces90] with smoothing coefficient ('m') set to unity. Formally,

$$p(v = u \mid c) \;\widehat{=}\; \frac{\#(C \rhd \{c' : \mathrm{Case} \mid (v, u) \in c' \wedge c \subseteq c'\}) + k}{\#(C \rhd \{c' : \mathrm{Case} \mid v \in \mathrm{dom}\, c' \wedge c \subseteq c'\}) + 1} \tag{3.8}$$

$$\underline{\textit{where}} \quad k \;\widehat{=}\; \begin{cases} p(v = u \mid \{\}) & c \neq \{\} \\ \dfrac{1}{\#\Theta(\!|v|\!)} & c = \{\} \end{cases}$$

It is easily shown by induction on the size of the set Var that I_B is necessarily a a valid probability distribution. Notice also that if the training sample is empty, then no information is obtained about the prior distribution.

Lemma 8

$$C = \langle\rangle \vdash I_B = \emptyset$$

The resulting program $\mathcal{D}(I_B)$ is then the identity function (Lemma 3).

3.2 Concrete Programs

Let us now refine abstract diagnostic programs to a form which is more feasibly implementable. For the purpose of our comparative study of diagnostic accuracy, we set one task: determination of the final diagnosis. The final diagnosis represents the definitive cause of the pain, and in our study, possible causes are mutually exclusive. The final diagnosis is therefore recorded in a single variable which we denote Δ.

$$\Delta : \mathrm{Var} \tag{3.9}$$

The possible values for Δ form the set of all possible final diagnoses. For convenience we refer to members of this set as *diseases*, although not all are strictly diseases (e.g. non-specific pain).

$$\mathrm{Disease} \mathrel{\widehat=} \theta(\!|\Delta|\!) \tag{3.10}$$

Our abstract model (DP) of diagnostic programs regards input patient descriptions as joint distributions over total cases. This allows input patient descriptions to be arbitrarily complicated specifications; recall $\mathcal{K}(s)$ represents the information conveyed by the assertion that specification s is met (Equation 2.10). Furthermore doubt about the presence or absence of findings can even be quantified probabilistically. However, in our present study, whenever a variable is recorded it is assigned just a single possible value. Our patient descriptions are therefore partial cases. The form of diagnostic program therefore we wish to implement is one that takes as input a partial case and returns a probability distribution over diseases. We refer to such a program as a *concrete program* (CP).

$$\mathrm{CP} \mathrel{\widehat=} \mathrm{Case} \nrightarrow \mathbb{D}\,\mathrm{Disease} \tag{3.11}$$

A description of a patient as a partial case c is equivalent to an assertion that the patient meets specification $\mathcal{S}(c)$ (all possible total reconstructions) where

$$\mathcal{S} : \mathrm{Case} \to \mathrm{Spec} \tag{3.12}$$

$$\forall c : \mathrm{Case} \bullet \mathcal{S}(c) = \{c' : \mathrm{CASE} \mid c \subseteq c'\}$$

A patient description as a partial case c therefore conveys information $\mathcal{K}(\mathcal{S}(c))$. This is the input to an abstract diagnostic program. If the program then outputs information I, this implies marginal disease distribution $\mathcal{M}(I)$ where

$$\mathcal{M} : \mathrm{Info} \to \mathbb{D}\,\mathrm{Disease} \tag{3.13}$$

$$\forall I : \mathrm{Info} \bullet \mathcal{M}(I) = \lambda d : \mathrm{Disease} \bullet \sum_{c : \mathrm{CASE} \mid c(\Delta) = d} I(c)$$

Therefore any abstract diagnostic program D can be implemented as concrete program $\mathcal{C}(D)$ where

$$\mathcal{C} : \mathrm{DP} \to \mathrm{CP} \tag{3.14}$$

$$\forall D : \mathrm{DP} \bullet \mathcal{C}(D) = \mathcal{S} \fatsemi \mathcal{K} \fatsemi D \fatsemi \mathcal{M}$$

We reserve the symbol ψ for concrete programs. The concrete program based on a Bayesian network is

$$\psi_B \;\widehat{=}\; \mathcal{C}(\mathcal{D}(I_B)) \tag{3.15}$$

This concrete program is easily implemented using the algorithm described by Lauritzen and Spiegelhalter for computing conditional probabilities across Bayesian networks [Lau88]. Notice that if the input partial case description happens to include the final diagnosis too, then the output is the corresponding delta distribution.

Lemma 9

$$c : \mathrm{Case} \mid \Delta \in \mathrm{dom}\, c \;\vdash\; \psi_B(c) = \lambda d : \mathrm{Disease} \bullet \begin{cases} 0 & d \neq c(\Delta) \\ 1 & d = c(\Delta) \end{cases}$$

(This result holds for *any* concrete program constructed from an abstract one, not just those based on Bayesian networks.)

3.2.1 Independence Bayes

A special class of Bayesian network is one in which Δ has no parents, and all other variables have only Δ as their parent. This is often referred to as the independence model. It has parents relation

$$P_I \;\widehat{=}\; \{v : \mathrm{Var} \mid v \neq \Delta \bullet v \mapsto \Delta\} \tag{3.16}$$

The corresponding concrete program has a familiar definition, usually called the *independence Bayes* formula.

Lemma 10

$$c : \mathrm{Case} \mid \Delta \notin \mathrm{dom}\, c \wedge P = P_I \;\vdash\; \psi_B(c) = \lambda d : \mathrm{Disease} \bullet \frac{p(\Delta = d \mid \{\}) \prod_{v:\mathrm{dom}\, c} p(v = c(v) \mid \{\Delta \mapsto d\})}{\sum_{d':\mathrm{Disease}} p(\Delta = d' \mid \{\}) \prod_{v:\mathrm{dom}\, c} p(v = c(v) \mid \{\Delta \mapsto d'\})}$$

This program is particularly simple to implement as the above formula is easily computed.

3.3 Exemplar Models

Consider the factors which determine our expectation of the relative frequency of cases in the actual population. Four essentially different factors can be distinguished.

1. The structure and function of the body, and its response to disease. (Anatomy, physiology and pathology.)
2. Incompleteness of our knowledge of the above. (Patient idiosyncrasy.)
3. The tendency for clinicians to differ in their history-taking and examination. (Observer error.)

4. The tendency for a patient's recollection of past events to alter and fade with time. (Long-term persistence of data.)

The Bayesian network representation encourages us to model all four factors collectively. (They can be separated only by introducing further variables.) However, the first factor represents deeper knowledge of a kind elicited only by specific experiments. This of course may still be probabilistic in nature: for example, the prevalence of a particular kind of anatomical variation. Nevertheless, representation of medical knowledge (the first factor) is simpler if it can be separated from other extraneous factors. Therefore an alternative approach to that of the Bayesian network is to provide an idealized description (J_E) of the population based on background knowledge, and combine this with a smoothing function Z (representing the other three factors collectively) which randomly transforms any theoretical case into an actual case. The information conveyed by the pair (J_E, Z) is given by the convolution

$$I_E \mathrel{\widehat=} \lambda c' : \textsc{Case} \bullet \sum_{c:\textsc{Case}} J_E(c) Z(c, c') \tag{3.17}$$

Perhaps the most rudimentary way of specifying J_E is to provide a prototypical example $E(d)$ of each disease d.

$$E : \text{Disease} \rightarrow \textsc{Case} \tag{3.18}$$

$$\forall d : \text{Disease} \bullet E(d)(\Delta) = d$$

The theoretical distribution J_E is then very sparsely populated. Each exemplar in the range of E is associated with the prior probability of the corresponding disease, and all other cases have zero probability.

$$J_E(c) \mathrel{\widehat=} \begin{cases} 0 & c \neq E(c(\Delta)) \\ p(\Delta = c(\Delta) \mid \{\}) & c = E(c(\Delta)) \end{cases} \tag{3.19}$$

3.3.1 A Smoothing Function

The smoothing function Z randomly modifies all observations. A simple model for this is to assume that all observations modify their values independently, while the disease remains fixed. This implies the following definition.

$$Z(c, c') \mathrel{\widehat=} \begin{cases} 0 & c'(\Delta) \neq c(\Delta) \\ \prod_{v:\text{Var} \mid v \neq \Delta} p(v : c(v) \rightsquigarrow c'(v)) & c'(\Delta) = c(\Delta) \end{cases} \tag{3.20}$$

Here the term $p(v : u \rightsquigarrow u')$ stands for the conditional probability that variable v is actually observed to have value u' given that the theoretically predicted value is u, for any v, u and u'. This is estimated from the training sample C by a modified form of Equation 3.8.

$$p(v : u \rightsquigarrow u') \mathrel{\widehat=} \tag{3.21}$$

$$\frac{\#(C \upharpoonright \{c : \text{Case} \mid (v, u) \in c \wedge \Delta \in \text{dom}\, c \wedge E(c(\Delta))(v) = u\}) + p(v = u' \mid \{\})}{\#(C \upharpoonright \{c : \text{Case} \mid v \in \text{dom}\, c \wedge \Delta \in \text{dom}\, c \wedge E(c(\Delta))(v) = u\}) + 1}$$

The concrete program based on the exemplar method is

$$\psi_E \triangleq \mathcal{C}(\mathcal{D}(I_E)) \tag{3.22}$$

This has a formula similar to that for independence Bayes.

Lemma 11

$c : \text{Case} \mid \Delta \notin \text{dom}\, c \vdash$

$$\psi_E(c) = \lambda d : \text{Disease} \bullet \frac{p(\Delta = d \mid \{\}) \prod_{v:\text{dom}\, c} p(v : E(d)(v) \rightsquigarrow c(v))}{\sum_{d':\text{Disease}} p(\Delta = d' \mid \{\}) \prod_{v:\text{dom}\, c} p(v : E(d')(v) \rightsquigarrow c(v))}$$

Notice that in general this requires fewer parameters to be estimated from the training data than are required by independence Bayes. This is because the predicted values of a variable v may be the same given two different diseases, d_1 and d_2.

$$E(d_1)(v) = E(d_2)(v)$$

so for any case c,

$$p(v : E(d_1)(v) \rightsquigarrow c(v)) = p(v : E(d_2)(v) \rightsquigarrow c(v))$$

Thus a single parameter is estimated from pooled training cases that were previously partitioned in order to estimate the two parameters that independence Bayes requires: $p(v = E(d_1)(v) \mid c(v))$ and $p(v = E(d_1)(v) \mid c(v))$. The exemplar model therefore exploits background knowledge to make more efficient use of training data.

Chapter 4

Filtering methods

Some methods are best described directly as concrete programs, rather than as full distribution transformers. This chapter describes a class of these methods, which all work by filtering the sequence of training cases to a subsequence relevant to the case to be diagnosed.

4.1 Filters

The diagnostic methods described in previous chapters all define joint probability distributions over cases, from which one can extract the information necessary to make a diagnosis for any particular case. Some methods, however, are best understood not as calculating a joint distribution but as directly calculating marginal disease probabilities. Thus, these methods should be specified as concrete programs (CP) rather than as diagnostic programs (DP).

One kind of method is to *filter* the sequence of training cases to a subsequence relevant to the case to be diagnosed, then estimate the probabilities of diseases from their frequencies in this subsequence. We shall represent a subsequence of the training cases by the *indices* of its members in the full sequence.

$$\text{Indices} \;\widehat{=}\; \mathbb{P}\,\text{dom}\,C \tag{4.1}$$

A *filter* is a function which maps any case to a subsequence of the training cases.

$$\text{Filter} \;\widehat{=}\; \text{Case} \to \text{Indices} \tag{4.2}$$

Any subsequence (represented as a set of indices, s) defines a disease distribution, using the Cestnik formula to estimate probabilities from frequencies.

$$dist(s) \;\widehat{=}\; \lambda d : \text{Disease} \bullet \frac{\#\{i : s \mid (\Delta, d) \in C[i]\} + 1/\#\text{Disease}}{\#\{i : s \mid \Delta \in \text{dom}\,C[i]\} + 1} \tag{4.3}$$

Clearly, this function is closely related to the function p (Equation 3.8) defined in Chapter 3. A concrete program is obtained by composing a filter with *dist*.

4.1.1 Some Examples

The null filter always returns the empty sequence.

$$null \mathrel{\widehat=} \lambda c : \mathrm{Case} \bullet \{\} \tag{4.4}$$

The null filter defines the concrete program which always returns the uniform distribution over diseases.

Lemma 12

$$\vdash null \fatsemi dist = \lambda c : \mathrm{Case} \bullet \lambda d : \mathrm{Disease} \bullet \frac{1}{\#\mathrm{Disease}}$$

Another simple filter is that which never restricts the training cases; the corresponding concrete program always returns the prior marginal distribution over diseases.

$$prior \mathrel{\widehat=} \lambda c : \mathrm{Case} \bullet \mathrm{dom}\, C \tag{4.5}$$

Lemma 13

$$\vdash prior \fatsemi dist = \lambda c : \mathrm{Case} \bullet \lambda d : \mathrm{Disease} \bullet p(\Delta = d \mid \{\})$$

A more useful filter is that which selects only those training cases that match the case to be diagnosed on all observations.

$$exact \mathrel{\widehat=} \lambda c : \mathrm{Case} \bullet \{i : \mathrm{dom}\, C \mid c \subseteq C[i]\} \tag{4.6}$$

This exact match filter defines a concrete program in a similar way to the simple method described in Section 2.4. Just as for that earlier example, this concrete program is also too naive; it can assist the diagnosis of a patient only if there is some previous case in C which meets the new patient's specification. This difficulty can be overcome by relaxing the requirement that the relevant cases must match the new case exactly; instead, the relevant cases are taken to be those which are 'sufficiently similar' to the new case. Different measures of similarity lead to different filters.

4.2 Nearest Neighbours

One approach is to define similarity in terms of 'closeness' under some distance function δ between cases.

$$\delta : \mathrm{Case} \times \mathrm{Case} \to \mathbf{R} \tag{4.7}$$

Such a function gives a measure of the dissimilarity of any two cases; the greater their dissimilarity, the greater the returned value. Given such a distance function and some $k \in \mathbf{N}$, the *nearest neighbours* method selects the k training cases closest to the case to be diagnosed. Formally, the nearest neighbours filter nn is defined by

$$nn : \mathrm{Filter} \tag{4.8}$$

$$\forall c : \mathrm{Case} \bullet \#nn(c) = k$$
$$\forall i, j : \mathrm{dom}\, C \mid \delta(c, C[i]) < \delta(c, C[j]) \bullet j \in nn(c) \Rightarrow i \in nn(c)$$

The filter has been deliberately underspecified, to the extent that no policy is given for breaking a tie for the k^{th} nearest neighbour. All that is required is that no other training case *is* closer to the new patient than the k cases selected. The corresponding concrete program is defined

$$\psi_K \;\widehat{=}\; nn_{\S}\mathbf{D} \tag{4.9}$$

Provided that the distance between any two cases can be feasibly computed, the nearest neighbours method is readily implemented. Although diagnosis of each new case requires comparison with every previous case, the number of computational steps grows only linearly with respect to the size of the training set.

4.2.1 Hamming Distance

There are many candidates for the distance function δ. Since our cases are described by categorical variables, the *Hamming distance* is one appropriate metric. This defines the distance between two cases to be the number of variables on which they differ. This informal definition must be refined for our purposes, since variables can be unrecorded in our cases. This refinement can be made in several ways.

The first is an 'optimistic' approach, where an unrecorded variable is assumed to match any value. Thus, only those variables that are recorded in both cases and on which the cases differ count towards the distance. Assuming that $\mathbf{N} \subseteq \mathbf{R}$, this (pseudo)metric is defined formally as

$$\delta_o \;\widehat{=}\; \lambda c, d : \text{Case} \bullet \#(\text{ran}\, c \cap \text{ran}\, d) - \#(c \cap d) \tag{4.10}$$

This function is not a true metric since distinct cases can be zero distance apart. In particular the empty case, in which no variables are recorded, is zero distance from any other case so will always be one of the nearest neighbours. Hence in practice this metric is unlikely to perform well since it favours cases for which many variables are unrecorded.

A second approach takes the pessimistic view that an unrecorded variable never matches any value. Thus a variable counts towards the distance between two cases *except* when the two cases agree on the known value of that variable. This metric is defined as

$$\delta_p \;\widehat{=}\; \lambda c, d : \text{Case} \bullet \#\text{Var} - \#(c \cap d) \tag{4.11}$$

Notice that this function is also not a true metric, since identical cases are only zero distance apart if they have no unrecorded variables.

Alternatively regard a case as a bit-vector, with one bit for each possible fact; if a fact *is* present its bit is on, otherwise it is off. The Hamming distance between two cases *in* this representation is a compromise between δ_o and δ_p. A variable unrecorded in one of the cases counts one towards the distance since one bit mismatches; a variable recorded in both cases but with different values counts two towards the distance, since two bits mismatch. This metric is described formally by

$$\delta_f \;\widehat{=}\; \lambda c, d : \text{Case} \bullet \#(c \cup d) - \#(c \cap d) \tag{4.12}$$

4.2.2 Bayesian Metric

Hamming distance suffers from a defect as a metric; the symptoms and signs recorded in a patient's case history are not sufficiently abstract for genuine similarities between cases to

be revealed. Typically only some of the observations concerning a particular patient will be significant indicants for or against diseases. With the Hamming metric, these observations can be masked by variation among the variables that are irrelevant for that patient. This problem can be circumvented by first mapping cases to a more abstract representation, and then applying the nearest neighbours method in this new space.

Clearly, the crucial step is finding a suitable abstract representation of cases. The ideal is that the representation should capture precisely those details that are diagnostically relevant. The most abstract possible representation, for the purposes of diagnosis, is the actual disease that the patient has. Since the aim is to discover the diagnosis, the actual disease will not be known. However, an estimate of the disease probabilities *can* be made using one of the concrete programs derived so far. Thus, we can represent a case by the disease probabilities calculated for it by some concrete program. The distance between two cases is then taken to be the Euclidean distance between their abstract representations; Euclidean distance *is* generally regarded as an optimal metric for nearest neighbours [Tod89, Sal91]. For example, using the program ψ_B derived from the Bayesian Network classifier, we get the following (pseudo)metric:

$$\delta_b \;\widehat{=}\; \lambda c, d : \mathrm{Case} \bullet \sum_{u:\mathrm{Disease}} (\psi_B(c)(u) - \psi_B(d)(u))^2 \tag{4.13}$$

This is not a true metric either, unless ψ_B is injective, since two distinct cases will be zero distance apart if they have identical estimated disease distributions.

4.3 Iterative Partitioning

There are other ways of defining similarity between cases besides the nearest neighbours approach. A case, formally defined as a function from Var to Val, can also be regarded as a collection of *facts*, each specifying the value of one of the variables. Formally, a fact is a pairing of a variable and a value.

$$\mathrm{Fact} \;\widehat{=}\; \mathrm{Var} \times \mathrm{Val} \tag{4.14}$$

The naive filter of definition 4.6 asserts that a training case is sufficiently similar to the case to be diagnosed only if it matches on all known facts. If this is too stringent a requirement, it can be relaxed by needing only that training cases must match some subset of the facts in the case to be diagnosed.

Any fact defines a subsequence of the training cases, namely, precisely those cases in which the given fact is true. The function *match* returns this subsequence.

$$match : \mathrm{Fact} \rightarrow \mathrm{Indices} \tag{4.15}$$

$$match \;\widehat{=}\; \lambda f : \mathrm{Fact} \bullet \{i : \mathrm{dom}\, C \mid f \in C[i]\}$$

This function uses a fact to partition the training cases; those which match the fact are kept and those which do not are discarded. Given a case to diagnose, the 'best' fact on which to partition can be selected (according to some criterion), and the sequence of training cases restricted using *match*. This partitioning process can be applied iteratively, repeatedly selecting the best fact for partitioning the current subsequence, until some stopping criterion

is met. We call this method of restricting the training cases to a relevant subsequence, *iterative partitioning*.

Selecting a succession of facts yields a sequence of ever-smaller subsequences of the training cases. We shall refer to such a decreasing sequence as a *chain* of subsequences. The training sequence is unrestricted if no facts are selected; thus the first element of the chain is the full training sequence. The last element in the chain is the subsequence of relevant cases we desire. This is captured formally by the function *iter*.

$$iter : \text{Filter} \tag{4.16}$$

$$\forall c : \text{Case} \bullet \exists S : \text{seq Indices} \bullet (\text{dom } C) \frown S \frown (iter(c)) \ \underline{chain}\ c$$

The relation *chain* specifies how to choose the facts on which to restrict the training sequence; there are several criteria which this relation should meet.

Clearly, we must ensure that the chain of subsequences *is* formed by successive restriction to those cases that match some fact. This property of sequences of indices is captured by the relation *decr*.

$$\_\ \underline{decr}\ \_ : \text{seq Indices} \leftrightarrow \text{Case} \tag{4.17}$$

$$S\ \underline{decr}\ c$$
$$\Leftrightarrow$$
$$\forall i : 1 \mathinner{..} \#S - 1 \bullet \exists f : c \bullet S[i+1] = S[i] \cap match(f)$$

To decide which fact to choose, we need some measure of the worth of restricting on any given fact. Let us suppose therefore that we have a *significance function*, σ, which when given a subsequence of the training cases and some proposed restriction of that subsequence, indicates whether the restriction is worthwhile. The definition is parametrised by a threshold α; a score above this threshold indicates that the proposed restriction is not worthwhile. The only axiom is that the vacuous restriction, which doesn't actually restrict the training cases, is never considered worthwhile.

$$\sigma : \text{Indices} \times \text{Indices} \rightarrow \mathbf{R} \tag{4.18}$$

$$\forall s : \text{Indices}; \bullet \sigma(s, s) > \alpha$$

We can now specify a second requirement of the chain of subsequences; each restriction is the best possible under the circumstances; no fact can be found which yields a better restriction. This property is captured by the relation *maximal*.

$$\_\ \underline{maximal}\ \_ : \text{seq Indices} \leftrightarrow \text{Case} \tag{4.19}$$

$$S\ \underline{maximal}\ c$$
$$\Leftrightarrow$$
$$\forall i : 1 \mathinner{..} \#S - 1; f : c \bullet \sigma(S[i], S[i] \cap match(f)) \geq \sigma(S[i], S[i+1])$$

Finally, at each stage the subsequence should be further restricted if and only if there is some fact (in the case to diagnose) which yields a worthwhile restriction. This property

dictates when partitioning should stop and is captured by the relation *signif*.

$$\_ \ \underline{signif} \ \_ : \mathbf{seq}\,\mathrm{Indices} \leftrightarrow \mathrm{Case} \tag{4.20}$$

$$S \ \underline{signif} \ c$$
$$\Leftrightarrow$$
$$\forall i : 1 \,..\, \#S \bullet i = \#S \Leftrightarrow \forall f : c \bullet \sigma(S[i], S[i] \cap match(f)) > \alpha$$

The relation *chain* is defined to meet all three requirements; it is the intersection of the three relations defined above.

$$\_ \ \underline{chain} \ \_ : \mathbf{seq}\,\mathrm{Indices} \leftrightarrow \mathrm{Case} \tag{4.21}$$

$$\underline{chain} \ = \ \underline{decr} \ \cap \ \underline{signif} \ \cap \ \underline{maximal}$$

A reasonably simple choice of significance function σ is the likelihood ratio of obtaining the disease frequencies observed in the restricted sample, under the assumption that the original sample provides the disease probabilities. Formally,

$$\sigma(s,t) \ \widehat{=} \ \#t^{\#t} \prod_{u:\mathrm{Disease}} \frac{dist(s)(u)^{n_u}}{n_u^{n_u}} \tag{4.22}$$
$$\underline{where} \quad n_u = \#(t \cap match(\Delta \mapsto u))$$

The concrete program implementing iterative partitioning is thus defined by

$$\psi_I \ \widehat{=} \ iter_{\S}\,\mathbf{D} \tag{4.23}$$

This is computationally more expensive than nearest neighbours because disease frequencies have to be recomputed at each iteration. Counting disease frequencies is linear in the size of the filtered sample, but in the worst case the filtered sample decreases in size by only one case at each iteration. The computational complexity is therefore quadratic in the size of the original training sample (although in practice very few iterations tend to be required).

Chapter 5

Neural Networks

This chapter presents a specification of a neural network approach to the design of a concrete program. Unlike the filtering methods, the whole of the training set is used to determine the distribution over diseases, not just a specially selected subset.

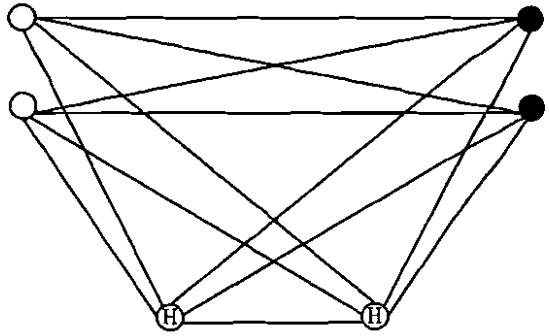
5.1 Feed-forward Networks

Over recent years, interest has revived in the use of networks of processing units for many purposes including signal processing, pattern recognition and classification [Rum86a, Rum86b, Lip87, Sim90]. Medical diagnosis has been one of the fields of application for these 'artificial neural networks' [Bou90, Low90, Gro90, Mul90, Bax91, Mac91, Aka92].

Although numerous different forms of network have been proposed, they do have common features. In general a network consists of a collection of processing nodes, linked by one-way weighted connections. Each node has some internal state, usually represented by a real number, and can transmit that value along the connections that lead from it. A node changes its value by forming the sum of the inputs it receives from other nodes, each scaled by the appropriate connection weight, then passing this through some 'activation function'. Networks are used by initialising the states of some or all of the nodes, then propagating these values around the network until it has reached a stable state. The final states of some or all of the nodes provide the result.

The most widely-used form of network is the *feed-forward perceptron*, in which there are no cycles of connections between nodes. There are three sets of nodes: the *input* nodes are the only nodes which are initialised, and they have no connections from any other node, and do not change state. The *output* nodes provide the result, and there are no connections from these nodes to any others. Between the input and output nodes are some *hidden* nodes. In the most general form of the feed-forward network, each output node has a connection from every input and hidden node, and the hidden nodes form a chain in which each has a connection from every input node and from those hidden nodes earlier in the chain. This is illustrated in Figure 5.1, which shows a network with two input, two hidden and two output nodes.

Figure 5.1: A feed-forward network. The lines represent connections from left to right.



○ Input node Ⓗ Hidden node ● Output node

5.1.1 Architecture

We wish to use a network of this form to define a concrete program. Since we require an estimate of probability for each disease we shall take a network with as many output neurons as there are diseases. With n hidden nodes, and the input nodes as yet unspecified, the set of all nodes is

$$\begin{aligned} \mathrm{Node} \;\widehat{=}\;\; & \mathrm{inp}\langle\!\langle \mathrm{Input} \rangle\!\rangle \\ \mid\; & \mathrm{hid}\langle\!\langle 1 \,..\, n \rangle\!\rangle \\ \mid\; & \mathrm{out}\langle\!\langle \mathrm{Disease} \rangle\!\rangle \end{aligned} \tag{5.1}$$

The state of all the nodes can he represented by a vector mapping each node to some real number.

$$\mathrm{Vector} \;\widehat{=}\; \mathrm{Node} \rightarrow \mathbf{R} \tag{5.2}$$

Since the input nodes do not change state, they are not really behaving as 'artificial neurons'. We therefore identify the subset of nodes which *do* behave in a neuron-like manner, namely the hidden and output nodes.

$$\mathrm{Neuron} \;\widehat{=}\; \mathrm{ran\,hid} \cup \mathrm{ran\,out} \tag{5.3}$$

Each neuron can have a (one-way) connection from any node, and there is a *connection weight* associated with each of these links. A connection weight scales any value passing along the connection, thus the absence of a connection between two nodes can be regarded as a connection with zero weight. The connections to any node can therefore be represented by a Vector giving connection weights from *all* nodes. The network architecture is determined

by the connection weights for all the neurons, and these are given by function A.

$$A : \text{Neuron} \rightarrow \text{Vector} \tag{5.4}$$

$$\forall n : \text{Neuron}; d : \text{Disease} \bullet A(n)(\text{ont}(d)) = 0$$
$$\forall i, j : 1 \mathinner{..} n \mid i \leq j \bullet A(\text{hid}(i))(\text{hid}(j)) = 0$$

5.1.2 Applying the Network

On receiving a collection of scaled inputs, a neuron takes their sum and passes it through an *activation function* to return a new value for that neuron. Many activation functions have been used for artificial neurons, but that applied most often is a sigmoid that maps real numbers to the range $[0, 1]$.

$$\phi = \lambda x : \mathbf{R} \bullet \frac{1}{1 + e^{-x}} \tag{5.5}$$

Given a vector V representing the current state of the network, and a second vector W representing the connection weights on the links to some node, $N(V, W)$ computes the new value for that node.

$$N : \text{Vector} \times \text{Vector} \rightarrow \mathbf{R} \tag{5.6}$$

$$\forall V, W : \text{Vector} \bullet N(V, W) = \phi \left(\sum_{n:\text{Node}} V(n)W(n) \right)$$

For a network to define a concrete program, some method must be found for presenting a case to a network for diagnosis. In a feed-forward network only the input nodes can be initialised, so let the function *enc* define an *encoding* of a case as an initial pattern of values for the input nodes.

$$enc : \text{Case} \rightarrow \text{Input} \rightarrow \mathbf{R} \tag{5.7}$$

A case is diagnosed by first encoding it then propagating the resulting values through the network. This behaviour is captured formally by the function E.

$$E : \text{Case} \rightarrow \text{Vector} \tag{5.8}$$

$$\begin{aligned} &\forall c : \text{Case}; i : \text{Input}; n : \text{Neuron} \bullet \\ &\quad E(c)(\text{inp}(i)) = enc(c)(i) \\ &\quad E(c)(n) = N(E(c), A(n)) \end{aligned}$$

The function is well-defined, despite the apparent circularity in the second clause of its definition. That this circularity is harmless follows from the network architecture specified by function A. The value of the first hidden neuron is determined solely by the values of the input nodes, since there are no other connections to that neuron. The value of each successive hidden neuron is then determined by the input values and the values of the preceding hidden neurons. Finally, the values of the output neurons are determined by the input nodes and hidden neurons, since no output neuron has a connection to any other.

A concrete program must return a probability distribution over the disease values. The output neurons, however, return values independently in the range $[0, 1]$, and it is unlikely that they will sum to one. Therefore in general the output values need to be normalised to obtain a concrete program.

$$\psi_N \;\widehat{=}\; \lambda c : \text{Case} \bullet d : \text{Disease} \bullet \frac{E(c)(\text{out}(d))}{\sum_{d' : \text{Disease}} E(c)(\text{out}(d'))} \tag{5.9}$$

5.1.3 Encoding and Training

There are many possible encodings of a case to a format suitable for input to a feed-forward network. One of the simplest is to have an input for every possible fact. Those inputs that correspond to facts present in the case are set to one, and the remaining inputs are set to zero. There is usually a 'bias node' in feed-forward networks which has a constant value. This bias node is best regarded as a distinguished input node, since it has connections to every neuron. The set of inputs is therefore defined as

$$\text{Input} \;\widehat{=}\; \text{bias} \mid \text{fact}(\langle\!\langle \text{Fact} \rangle\!\rangle) \tag{5.10}$$

With this fact-orientated representation of cases, the encoding function is defined by

$$\begin{aligned} &\forall c : \text{Case}; f : \text{Fact} \bullet \\ &\quad enc(c)(\text{bias}) = 1 \\ &\quad enc(c)(fact(f)) = \begin{cases} 1 & f \in c \\ 0 & f \notin c \end{cases} \end{aligned} \tag{5.11}$$

The definition of the network architecture, given earlier by the function A, places no restriction on connection weights other than those which are set to zero. To define an accurate concrete program, the network must be 'trained' by optimizing the other connection weights. This requires iterative adaptation. The objective is to minimise an error measure on the values produced at the output nodes for the training cases. The best-known method of doing this is *back-propagation* [Rum86a] which minimises the mean squared difference between the desired and actual values produced by the output neurons.

Chapter 6

A Probabilistic Rule-Based System

In this chapter a probabilistic model for inference rules is presented. A rule-based system is regarded as a Bayesian network in which each conditional probability table is specified implicitly by a collection of weighted rules. The weight of each rule is given a logistic interpretation and obtained from a training sample by standard optimization methods. It is also pointed out that the representation is also suitable for causal knowledge, offering a considerable parameter reduction compared to explicit Bayesian networks.

6.1 An Inferential Chain Decomposition

In many applications, rule-based inferential knowledge has proved a successful foundation for building expert systems [Sho76, Dud79, Fox80, Goo85, Jac86, Won90]. However, rule-based representations have been repeatedly criticized for the way uncertainty is handled [Spi84, Hec86, Spi86, Hec88, Nea89, Dan92]. In this chapter we show how weighted inference rules can be given a sound probabilistic interpretation based on the Bayesian network method previously described. This is possible because *any* chain decomposition is valid when constructing a Bayesian network. It is usual to adopt a causal ordering for the chain decomposition only because this tends to lead to the sparsest 'parents' relation, and hence the smallest conditional probability tables. However, if we choose the reverse ordering, we then represent *inferential* knowledge rather than causal knowledge. Therefore let us assume that our chain decomposition corresponds to the order in which the values of variables are usually inferred. Observable variables (symptoms and signs etc.) are anterior. Pathophysiological states and diseases follow posteriorly.

Of course a consequence of adopting an inferential chain decomposition rather than a causal one is that numbers of parents tend to be much larger. This means that conditional probability tables become infeasibly large either to estimate directly from a training sample (too many empty cells) or to store in a computer. A solution is to adopt a parametric model for each table. If the chain decomposition were causal then a natural assumption to make would be that multiple causes produce their common effects independently; this is usually referred to as the 'Noisy OR-Gate' model [Coo89, Shw91]. However, since our chain de-

composition is inferential, we require a parametric model for the conditional probability of a cause given everything that is known about its effects.

6.1.1 A Logistic Model

A suitable model appears to be the logistic one [And82]. Suppose we have a set of binary random variables $\alpha, \beta_1, \beta_2, \ldots \beta_n$. According to the logistic model, if α (the child) depends logistically on $\beta_1, \beta_2, \ldots \beta_n$ (the parents) then the conditional log-odds have a linear form.

$$\ln \frac{p(\alpha = 1 \mid \beta_1, \beta_2, \ldots \beta_n)}{p(\alpha = 0 \mid \beta_1, \beta_2, \ldots \beta_n)} = k_0 + k_1\beta_1 + k_2\beta_2 + \ldots + k_n\beta_n \tag{6.1}$$

where the $k_0, k_1, \ldots k_n$ are real-valued constants (the logistic weights). Notice how only $n+1$ parameters are thus required to specify a table of 2^n conditional probabilities. Equation 6.1 is consistent with several families of distribution. These include conditional independence of $\beta_1, \beta_2, \ldots \beta_n$ given both states of α, and more generally, log-linear conditional distributions with equal interaction terms. The logistic model is also consistent with mutual exclusivity of the $\beta_1, \beta_2, \ldots \beta_n$. Because of its generality, the logistic form has been widely used for combining evidence and taking interactions into account [And82, Lud83, Spi84, Sey90]. Furthermore, the logistic form can be made even more general if we allow the terms on the right-hand side of Equation 6.1 to contain arbitrary Boolean expressions $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_m$ over the variables $\beta_1, \beta_2, \ldots \beta_n$. For example,

$$\varepsilon_2 \;\hat{=}\; \beta_1 \wedge (\beta_5 \vee \neg\beta_7)$$

where $\wedge$, $\vee$, and $\neg$ denote minimum, maximum and complement of their arguments, respectively. Let **true** be the constant expression that always has value 1. We can then multiply the bias weight k_0 in Equation 6.1 by **true** so that every term on the right-hand side has the same form: a weighted expression. Thus more generally,

$$\ln \frac{p(\alpha = 1 \mid \beta_1, \beta_2, \ldots \beta_n)}{p(\alpha = 0 \mid \beta_1, \beta_2, \ldots \beta_n)} = k_1\varepsilon_1 + k_2\varepsilon_2 + \ldots + k_m\varepsilon_m \tag{6.2}$$

Any table of finite conditional odds can be represented by an equation of this form because if necessary we can have a set of mutually exclusive expressions on the right-hand side of the equation, one for each possible state of the variables $\beta_1, \beta_2, \ldots \beta_n$. This though would mean that $m = 2^n$, so no saving would be achieved. Nevertheless, for practical purposes it should be possible to achieve a satisfactory approximation with a reasonable number of expressions selected in the light of expert knowledge of the field of application.

However, one of the several requirements of a probabilistic knowledge representation listed by Lauritzen and Spiegelhalter [Lau88] is that there should be no difficulty in handling total logical dependence between variables. Logical dependence would entail infinitely large conditional odds, and thus infinitely large logistic weights. Fortunately, this is easily circumvented by transforming the weights to the open interval $(0, 1)$ [Haj85]. A suitable transformation is $\mathcal{G}$ where

$$\mathcal{G} \;\hat{=}\; \lambda x : \mathbf{R} \bullet \frac{\exp(x)}{\exp(x) + 1} \tag{6.3}$$

Transformed weights combine not by simple addition but by a different operator $\oplus$. Thus, for any p and q such that $0 < p, q < 1$,

$$p \oplus q = \mathcal{G}(\mathcal{G}^{-1}(p) + \mathcal{G}^{-1}(q)) \tag{6.4}$$

Since $\mathcal{G}$ is bijective, the operator $\oplus$ inherits the properties of commutativity and associativity from simple addition in terms of which it is defined. The operator also has $\frac{1}{2}$ as its identity element. Furthermore, substituting for $\mathcal{G}$ in Equation 6.4 we obtain the following more convenient rule of combination. We take this as the actual definition of $\oplus$.

$$p \oplus q \;\widehat{=}\; \frac{pq}{pq + (1-p)(1-q)} \tag{6.5}$$

The extreme weights 0 and 1 are both zero elements of the operator $\oplus$. In the presence of complete certainty, further evidence makes no difference. These weights denote logical entailment.

Lemma 14

$$p : \Pr \mid p \neq 1 \vdash 0 \oplus p = 0 = p \oplus 0$$

Lemma 15

$$p : \Pr \mid p \neq 0 \vdash 1 \oplus p = 1 = p \oplus 1$$

Notice that the two zero elements cannot be combined together by $\oplus$ since this denotes logical contradiction; $0 \oplus 1$ and $1 \oplus 0$ are both undefined.

If we apply function $\mathcal{G}$ to both sides of Equation 6.2 we obtain

$$p(\alpha = 1 \mid \beta_1, \beta_2, \ldots \beta_n) = \bigoplus_{j:1..m \mid \varepsilon_j = 1} k'_j \tag{6.6}$$

where the k'_j represent the transformed weights ($k'_j = \mathcal{G}(k_j)$). If all expressions ε_j evaluate to zero, then the right-hand side of Equation 6.6 is simply the identity element of the $\oplus$ combinator ($\frac{1}{2}$). Any conditional probability table can be represented by an equation of the above form, even a table containing the extreme probabilities of 1 and 0.

6.1.2 Multi-Valued Variables

A difficulty still preventing the direct application of the logistic model is the requirement that all variables are binary. In most real domains such as medical diagnosis, variables have more than two possible values. For example, four values are required to describe the progress of a patient's pain: 'stopped', 'better', 'same' and 'worse'. A solution is to make the nodes of our Bayesian network *specifications* rather than variables. The Bayesian network thus defines a joint probability distribution of truth assignments to these specifications. Assignment of 'true' to a specification s means that s is met, and assignment of 'false' means that the complement $\bar{s}$ is met instead. Formally, let Q be a sequence of specifications.

$$Q : \mathrm{seq}\,\mathrm{Spec} \tag{6.7}$$

We assume that Q has an inferential ordering. Anterior specifications (those appearing early in Q) thus concern the presence of symptoms and signs etc. Satisfaction of these specifications is directly observable. Posterior specifications (those appearing later in Q) concern pathophysiological states and diseases. Satisfaction of these specifications is usually not observed directly, but instead inferred from the truth values of the anterior ones.

Any assignment of truth values to a set of specifications *itself* constitutes a specification, namely the conjunction (intersection) of all specifications assigned 'true' and the complements of all specifications assigned 'false'. Therefore in our formalism we will regard a truth assignment simply as a specification. For any set of specifications S, $\mathcal{A}(S)$ denotes the set of all truth assignments to the members of S.

$$\mathcal{A} : \mathbb{P}\,\mathrm{Spec} \rightarrow \mathbb{P}\,\mathrm{Spec} \tag{6.8}$$

$$\forall S : \mathbb{P}\,\mathrm{Spec} \bullet \mathcal{A}(S) = \{S' : \mathbb{P}\,S \bullet \bigcap S' - \bigcup(S - S')\}$$

Notice that $\mathcal{A}(S)$ is a partition of the universal specification.

Lemma 16

$$S : \mathbb{P}\,\mathrm{Spec} \vdash \bigcup \mathcal{A}(S) = \textsc{Case} \ \wedge \ \forall s, s' : \mathcal{A}(S) \mid s \neq s' \bullet s \cap s' = \{\}$$

Furthermore, if S is empty, only one truth assignment is possible, and this corresponds to the universal specification.

Lemma 17

$$\vdash \mathcal{A}(\{\}) = \{\mathrm{Case}\}$$

Our task is to specify for each i the conditional probability that the random case meets specification $Q[i]$ given all possible combinations of truth-value assignments to the anterior specifications $Q(1 \,..\, i-1)$. Let T be a sequence of conditional probability tables, one for each member of Q.

$$T : \mathrm{seq}\,\mathrm{Table} \tag{6.9}$$

$$\#T = \#Q$$

We represent each such table by a logistic form similar to that of Equation 6.6. Each expression (ε_j) is now a Boolean combination of anterior *specifications* rather than of binary variables. Each such expression itself constitutes a specification. It is a disjunction of truth assignments to those anterior specifications. For any set of specifications S, let $\mathcal{E}(S)$ denote the set of all 'expressions' that can be constructed.

$$\mathcal{E} : \mathbb{P}\,\mathrm{Spec} \rightarrow \mathbb{P}\,\mathrm{Spec} \tag{6.10}$$

$$\forall S : \mathbb{P}\,\mathrm{Spec} \bullet \mathcal{E}(S) = \{S' : \mathbb{P}\,\mathcal{A}(S) \bullet \bigcup S'\}$$

A table is thus a set of weighted expressions (specifications). Each expression can be regarded as the antecedent of an inference rule whose conclusion is the corresponding specification in sequence Q and whose certainty factor is the weight of the expression.

$$\mathrm{Table} \ \widehat{=} \ \mathrm{Spec} \nrightarrow \mathrm{Pr} \tag{6.11}$$

However, the expressions appearing in a table must be constructed only from anterior specifications. We call such a table *well-formed*. This guarantees that it is necessary to know only the truth assignment to anterior probabilities in order to evaluate the conditional probability of meeting $Q[i]$, for each i.

$$\forall i : 1 \mathinner{..} \#Q \bullet \operatorname{dom} T[i] \subseteq \mathcal{E}(Q(\!|1 \mathinner{..} i-1|\!)) \tag{6.12}$$

Any such table though must be *internally consistent*. Under no circumstances must it be possible to derive a logical contradiction, otherwise the corresponding conditional probability would be undefined.

$$\forall i : 1 \mathinner{..} \#Q; s, s' : \mathrm{Spec} \bullet \{s \mapsto 0, s' \mapsto 1\} \subseteq T[i] \Rightarrow s \cap s' = \{\} \tag{6.13}$$

Finally, for each i, if a truth assignment to anterior specifications logically determines whether or not $Q[i]$ is met, then table $T[i]$ should evaluate to the corresponding extreme probability of 1 or 0. This is ensured by including an appropriate expression in the table with extreme weight of 1 or 0, accordingly. If the sequence of tables T has this property, then we say it is *externally consistent* with respect to sequence Q.

$$\begin{aligned} \forall i : 1 \mathinner{..} \#Q \bullet \forall s : \mathcal{A}(Q(\!|1 \mathinner{..} i-1|\!)) \mid s \neq \{\} \bullet \\ s \subseteq Q[i] \Rightarrow \exists s' : \operatorname{dom} T[i] \bullet s \subseteq s' \wedge T[i]s' = 1 \\ s \subseteq \overline{Q[i]} \Rightarrow \exists s' : \operatorname{dom} T[i] \bullet s \subseteq s' \wedge T[i]s' = 0 \end{aligned} \tag{6.14}$$

For any truth assignment s to the sequence of specifications Q, if s is non-empty ($s \neq \{\}$) then let $p_i(s)$ denote the conditional probability that $Q[i]$ is met given the truth assignments to the anterior specifications. This is determined by combining logistically the weights of all expressions in Table $T[i]$ which evaluate to 'true' (i.e. all rules which 'fire').

$$p_i(s) \mathrel{\hat{=}} \bigoplus_{s' : \operatorname{dom} T[i] \mid s \subseteq s'} T[i]s' \tag{6.15}$$

By taking the product of these conditional probabilities for all i, we obtain the joint probability $p(s)$.

$$p(s) \mathrel{\hat{=}} \prod_{i:1..\#Q} \begin{cases} p_i(s) & s \subseteq Q[i] \\ 1 - p_i(s) & s \subseteq \overline{Q[i]} \end{cases} \tag{6.16}$$

Thus Q and T together define a joint probability distribution over truth assignments to Q. If every non-empty truth assignment s is a singleton specification then Q is said to be *complete*. If Q is complete then the probability associated with any case c is given by the probability $p(\{c\})$ associated with the corresponding singleton specification. If not, then all members of s are taken to be equiprobable (Principle of Indifference).

The pair (Q, T) constitutes a knowledge base representing information I_R where

$$I_R \mathrel{\hat{=}} \lambda c : \textsc{Case} \bullet \frac{p(s)}{\#s} \tag{6.17}$$

$$\underline{\textit{where}} \quad s : \mathcal{A}(\operatorname{ran} Q) \mid c \in s$$

Notice that if the knowledge base is empty, then it conveys no information at all.

Lemma 18

$$Q = \langle\rangle \vdash I_R = \emptyset$$

The concrete program provided by the rule-based system is defined by

$$\psi_R \triangleq C(\mathcal{D}(I_R)) \tag{6.18}$$

6.1.3 Implementation

For simplicity, we restrict the conclusions of rules to atomic propositions. An *atomic proposition* 'v in U' is an assertion that the value of variable v lies in the set U. For any v and U there is a direct correspondence between the atomic proposition and an abstract specification.

$$v \text{ in } U \triangleq \{c : \textsc{Case} \mid c(v) \in U\} \tag{6.19}$$

We place no restriction on the antecedents of rules. Any specification can be represented as a Boolean expression over atomic propositions in which conjunction is intersection (s and $t \triangleq s \cap t$), disjunction is union (s or $t \triangleq s \cup t$), and negation is complementation (not $s \triangleq \overline{s}$). This provides a convenient shorthand for writing knowledge bases.

The choice of sequence Q, and the choice of inference rules (expressions appearing in tables T) are made in the light of expert knowledge of the domain of application. The weights associated with the expressions in T can be derived from the training sample C. Since for each i, table $T[i]$ comprises a logistic discriminant function for the truth value of specification $Q[i]$, standard iterative methods for fitting logistic discriminant functions can be applied. Each table can be treated independently of the others. The method we have implemented is iterative maximum likelihood estimation using mixture sampling with initial certainty factors equal to the identity ($\frac{1}{2}$) as suggested by Anderson [And82]. For expediency we implemented only simple gradient descent (gain = 1) rather than the more complicated quasi-Newton method. We adapt the certainty factors only after each pass through the entire training sequence. Unlike the neural network, this is feasible for the rule-based system because convergence is much faster. This is because there are only a few weights in each family of rules, whereas the neural network has many more parameters to optimize.

If the sequence Q has an inferential ordering, then observations made of any case c will tend to correspond to truth assignments to an initial segment of the chain. This allows statistical inference (calculation of $\psi_R(c)(d)$ for each disease d) to be made by Monte Carlo simulation along the sequence Q [Cor86]. Starting with the first proposition $Q[i]$ whose truth value is unknown, we compute the probability of $Q[i]$ being true given the anterior evidence (Equation 6.15), and the assign $Q[i]$ value true randomly with this probability, otherwise false. This procedure is then repeated for the next proposition $Q[i+1]$ in the sequence, and so on until all propositions in Q have been assigned a truth value. This samples the joint distribution of the unknown propositions conditional on the observed values; we call this a simulation *run*. We perform repeated simulation runs (e.g. 10,000) and count the relative frequency with which the disease variable (Δ) takes each possible value. This is an estimate of the distribution of Δ given the observations c.

Although we have developed this rule-based representation specifically for inferential knowledge, there is no formal requirement that Q has an inferential ordering. Q can just as easily

be given a causal ordering. Rules now relate causes to their effects rather than the other way around, but the rule-based representation still offers considerable parameter reduction over explicit representation of conditional probability tables. Although the 'Noisy OR-Gate' is usually employed for this purpose, the logistic model has the advantage of being able to handle inhibitory influences as well as causal ones, and in any case can usually approximate the 'Noisy OR-Gate' when required in practice [Tod93b]. Families of logistic rules thus form a flexible, readable and efficient representation both for inferential and causal knowledge.

If a causal ordering is chosen then the Monte Carlo simulation method is no longer suitable for drawing statistical inferences. Provided that the underlying Bayesian network on binary propositions is sufficiently sparse, the Lauritzen-Spiegelhalter algorithm [Lau88] offers a solution. If not, then a simple alternative is to generate a large sample of random cases from the model, and use this as training sequence C for some suitably flexible statistical classifier. If the model is correct, then the generated sample will be statistically indistinguishable from a sample of real cases. However, because it is generated from a computer model rather than laboriously collected by hand, it can be very much larger (e.g. 10^6 cases) and consequently convey more information.

Chapter 7

A Categorical Knowledge-Based System

In this chapter we present a formal model of a diagnostic program which employs categorical knowledge. Diagnostic inference is modelled by idempotent, decreasing functions on specifications. This leads to an implementation as a flowchart with the unusual feature that inferences are made by repeatedly traversing the flowchart until the same path is eventually followed.

7.1 Categorical Reasoning

The knowledge-based methods described in previous chapters have been concerned with probabilistic reasoning under uncertainty. But to what extent is it necessary to represent uncertainty, with all its associated complexity? Can similarly accurate (or better) results be achieved with purely categorical knowledge? Although 'algorithmic diagnosis' has been much criticized (e.g. [Sho79, Mac78]), there is recent evidence that carefully designed flowcharts can be acceptably accurate [Fra91]. In this chapter we develop a model of categorical reasoning.

Let us begin by dropping quantification of non-determinism. A case history contains many facts of diagnostic value: symptoms and signs, personal details such as age and sex, results of blood and urine tests, and the results of any other investigations performed such as X-rays. Furthermore, the patient may be known to suffer from one or more diseases such as diabetes or peptic ulceration, and these facts too will be recorded in the case history. A case history thus constitutes a 'specification' that the patient is asserted to meet. Given such a case history, the diagnostic task is to find a causal explanation for the patient's symptoms, signs and test results. Explanations may involve identification of a single disease, or of several coexistent diseases. A more refined explanation may also include pathophysiological states the patient has, such as shock or septicaemia. Thus an explanation can also be considered to be a specification, one which the patient is inferred to meet. The diagnostic task is then to find, by application of some suitable inference procedure, the strongest specification that the patient meets. Once we drop quantification of uncertainty, the most abstract view of a diagnostic program thus becomes a function on specifications rather than on information. We refer to such a program as an *inference procedure*.

We ascribe two properties to inference procedures. Firstly, we have available no element in the set 'Info' to represent an illogical case history. That is why we excluded illogical case histories from the domain of abstract diagnostic programs (DP) which were thus defined as partial functions. However, the set 'Spec' *does* contain a distinguished element which represents any illogical case history: the impossible specification {}. Therefore, we are now able to define inference procedures as *total* functions.

Secondly, since an inference procedure engages in categorical reasoning, it should be complete in the sense that it draws the fullest possible conclusion from the initial data. Therefore applying the same inference procedure to its own conclusion should derive nothing new (idempotency). Formally, let IP denote the set of all inference procedures.

$$\mathrm{IP} \;\widehat{=}\; \{p : \mathrm{Spec} \rightarrow \mathrm{Spec} \mid p \fatsemi p = p\} \tag{7.1}$$

Every inference procedure p can also be regarded as an abstract diagnostic program $\mathcal{D}(p)$. The abstraction function $\mathcal{D}$ is distinguished from the other function with the same symbol (Equation 2.6) by its type. Recall that for any specification s, $\mathcal{K}(s)$ represents the information conveyed by the assertion that s is met (Equation 2.10). Also notice that $\mathcal{K}$ is injective. Therefore, we define $\mathcal{D}$ formally:

$$\mathcal{D} : \mathrm{IP} \rightarrow \mathrm{DP} \tag{7.2}$$

$$\forall p : \mathrm{IP} \bullet \mathcal{D}(p) = \mathcal{K}^{-1} \fatsemi p \fatsemi \mathcal{K}$$

7.1.1 Ordering Specifications

Given that specifications are sets of case descriptions, the subset relation provides a natural partial ordering for specifications. We interpret this as a 'stronger than' relation.

Example 2 *Suppose that we have only four possible total cases*

$$\textsc{Case} = \{C_1, C_2, C_3, C_4\}$$

and consider two specifications s and t defined to be

$$s \;\widehat{=}\; \{C_1, C_2, C_3\}$$
$$t \;\widehat{=}\; \{C_1, C_3\}$$

Specification s asserts that the complete patient description may be one of C_1, C_2 or C_3, whereas t asserts that only C_1 and C_3 are possible. Therefore t is stronger than s because it is more deterministic.

The impossible specification ({}) is the strongest of all: no patient can meet it, since it has been strengthened to absurdity. The universal specification (CASE) is the weakest: since it is met by all patients it can never tell us anything new.

Since there is an ordering on specifications, should inference procedures be monotonic? Should stronger premises necessarily lead to stronger conclusions? One of the motivations for the development of non-monotonic logic was the need to reason from incomplete evidence. Since our specifications allow us to describe patients incompletely, we should not insist that inference procedures are monotonic.

7.1.2 Inference Steps

Inference tasks are usually too complex to be carried out at a single step. For example, we might consider defining an inference procedure by means of a single look-up table, but in practice the table would be far too large. Instead, reasoning tends to follow a sequence of steps, each increasing our knowledge of the case, until we finally reach a conclusion. Therefore we regard an *inference step* (IS) as a decreasing function on specifications; as a specification decreases in size, the stronger it becomes.

$$\mathrm{IS} \;\widehat{=}\; \{p : \mathrm{Spec} \to \mathrm{Spec} \mid \forall s : \mathrm{Spec} \bullet p(s) \subseteq s\} \tag{7.3}$$

Notice that inference steps are closed under composition.

Lemma 19

$$p, q : \mathrm{IS} \vdash p \fatsemi q \in \mathrm{IS}$$

From this it follows that an inference step can be freely repeated. Since IS is finite, this repetition eventually leads to a fixed point, and furthermore this fixed point is idempotent. Therefore this iterative method yields an inference procedure from any inference step. (Inference steps are analogous to UNITY [Cha88] programs, with specifications corresponding to program states.)

This is easily shown. Firstly, we promote the subset order on specifications to give a partial order on inference steps.

$$\_ \sqsubseteq \_ : \mathrm{IS} \leftrightarrow \mathrm{IS} \tag{7.4}$$

$$\forall p, q : \mathrm{IS} \bullet \\ p \sqsubseteq q \Leftrightarrow \forall s : \mathrm{Spec} \bullet p(s) \subseteq q(s)$$

Informally, $p \sqsubseteq q$ means that p is stronger than q; for any specification s, $p(s)$ is at least as deterministic as $q(s)$. Notice that composition strengthens an inference step.

Lemma 20

$$p, q : \mathrm{IS} \vdash p \fatsemi q \sqsubseteq p$$

Clearly the sequence of iterations p^n is a descending chain in the partial order. Since IS is finite it is well-founded with respect to the partial order, so the chain of p^n must reach a fixed point, which is the greatest lower bound of the chain. The function $\mathcal{F}$ finds this lower bound. (p^n denotes the composition of n copies of p.)

$$\mathcal{F} : \mathrm{IS} \to \mathrm{IS} \tag{7.5}$$

$$\forall p : \mathrm{IS} \bullet \\ \mathcal{F}(p) = \bigsqcap \{n : \mathbb{N} \bullet p^n\}$$

It follows from the antisymmetry of the partial order that $\mathcal{F}(p)$ is idempotent and thus a member of IP.

Lemma 21

$$p : \mathrm{IS} \vdash \mathcal{F}(p) \in \mathrm{IP}$$

7.2 A Representation for Inference Steps

In practice, individual inference steps are still too complex to specify by explicit enumeration; a more concise representation is needed. Some steps are easy to describe because they are so simple. An example is a 'constant' step in which the new information provided is independent of the premise. Given any specification s, $\mathcal{C}(s)$ is the constant step which when given any specification t as premise concludes that both s and t are met.

$$\mathcal{C} : \mathrm{Spec} \to \mathrm{IS} \tag{7.6}$$

$$\begin{aligned} &\forall s : \mathrm{Spec} \bullet \\ &\quad \mathcal{C}(s) = \lambda t : \mathrm{Spec} \bullet s \cap t \end{aligned}$$

The constant inference step $\mathcal{C}(\{\})$ simply rejects all case descriptions as impossible.

Lemma 22

$$\vdash \mathcal{C}(\{\}) = \lambda s : \mathrm{Spec} \bullet \{\}$$

Conversely, the constant inference step $\mathcal{C}(\textsc{Case})$ always returns its input unaltered, having deduced nothing new.

Lemma 23

$$\vdash \mathcal{C}(\textsc{Case}) = \lambda s : \mathrm{Spec} \bullet s$$

Just as simple assignments are insufficient for useful computer programming, constant inference steps are inadequate for useful reasoning. We need some other method of representing more powerful inference steps; one technique is to construct them from simpler steps, just as complex programs are composed from simple commands. We already have one combinator, sequential composition (Lemma 19). Unfortunately, this does not give us more power, since composition of two constant inference steps only yields another such step.

Lemma 24

$$s, t : \mathrm{Spec} \vdash \mathcal{C}(s) \fatsemi \mathcal{C}(t) = \mathcal{C}(s \cap t)$$

Another combinator used in programming languages is alternation. We can construct a conditional combinator for inferences steps too. Thus far, a specification has been used only as an assertion, but we can also regard it as a question asking *whether* that assertion is true. This means that specifications can be used as the guards for alternations. Given inference steps p, q and specification s, we define the construction $p \lhd s \rhd q$ to be an inference step that behaves like p if specification s is met, and like q otherwise.

$$\_ \lhd \_ \rhd \_ : (\mathrm{IS} \times \mathrm{Spec} \times \mathrm{IS}) \to \mathrm{IS} \tag{7.7}$$

$$\begin{aligned} &\forall p, q : \mathrm{IS}; s, t : \mathrm{Spec} \bullet \\ &\quad t \subseteq s \Rightarrow (p \lhd s \rhd q)(t) = p(t) \\ &\quad t \not\subseteq s \Rightarrow (p \lhd s \rhd q)(t) = q(t) \end{aligned}$$

The conditional combinator and function C are sufficient to represent any inference step as a tree where each node and leaf is labelled with a specification; a leaf specification represents a constant inference step, and a node specification represents the combination of the two inference steps that branch from it. Before defining these inference trees formally we need to look more closely at how they will handle missing data.

7.2.1 Missing Data

When carrying out medical diagnosis data are invariably missing; tests and X-rays may not have been carried out, and various signs may not have been recorded. It is important that a diagnostic method can always handle missing data, as Example 3 shows.

Example 3 *Suppose, having exhausted all other evidence, we have to decide which of two diagnoses, A and B, is the more probable on the basis of a Boolean indicant variable, v. Consider the situations described by the following two diagrams, which give the probabilities of (diagnosis,indicant) pairs.*

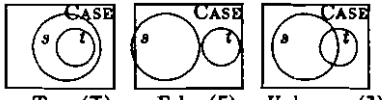
	$v=1$	$v=0$
A	0.3	0.1
B	0.2	0.4

	$v=1$	$v=0$
A	0.4	0.2
B	0.1	0.3

In both cases, if $v=1$ then we should diagnose A as the more probable, and if $v=0$ we should diagnose B. However, if the value of v is unknown, then the most probable diagnosis differs in the two situations. In the first case, B has the higher prior probability (0.6), whereas in the second case it is A that has the higher probability (0.6).

Clearly, given that a patient is known to meet specification t, the question 'Does the patient meet specification s?' has three possible answers: 'true', 'false' and 'unknown' (Figure 7.1).

Figure 7.1: Venn diagrams showing whether t meets s.



True (T) $t \subseteq s$ | False (F) $t \subseteq \overline{s}$ | Unknown (?) $t \not\subseteq s \wedge t \not\subseteq \overline{s}$

The last two cases are not distinguished by the alternation $p \lhd s \rhd q$; it behaves as q in both situations, and this is clearly inadequate as Example 3 shows. By using the combinator twice a three-way distinction can be made.

$$p \lhd s \rhd (q \lhd \overline{s} \rhd r)$$

Thus if the patient meets s the construction behaves as p, if instead the patient meets $\overline{s}$ it behaves as q, and otherwise it behaves as r.

7.2.2 Inference Trees

We therefore represent an inference step as a tree of specifications. The tree is ternary because each non-terminal node represents a question with three possible answers: true (T), false (F) and unknown (?).

$$\mathrm{IT} ::= \mathrm{leaf}(\langle\langle \mathrm{Spec} \rangle\rangle) \mid \mathrm{node}(\langle\langle \mathrm{Spec} \times \mathrm{IT} \times \mathrm{IT} \times \mathrm{IT} \rangle\rangle) \tag{7.8}$$

A semantic function $\mathcal{J}$ for inference trees can now be defined simply by applying the constant and alternation functions to the relevant parts of each tree.

$$\mathcal{J} : \mathrm{IT} \rightarrow \mathrm{IS} \tag{7.9}$$

$$\begin{aligned} &\forall T_{\mathsf{T}}, T_{\mathsf{F}}, T_{?} : \mathrm{IT}; s : \mathrm{Spec} \bullet \\ &\quad \mathcal{J}(\mathrm{leaf}(s)) = \mathcal{C}(s) \\ &\quad \mathcal{J}(\mathrm{node}(s, T_{\mathsf{T}}, T_{\mathsf{F}}, T_{?})) = \mathcal{J}(T_{\mathsf{T}}) \lhd s \rhd (\mathcal{J}(T_{\mathsf{F}}) \lhd \overline{s} \rhd \mathcal{J}(T_{?})) \end{aligned}$$

It follows by structural induction over IT that the result of applying $\mathcal{J}$ to any tree is a valid inference step. Furthermore, it can be shown by construction of a normal form tree that every inference step has a corresponding representation as an inference tree.

Lemma 25

$$\vdash \mathrm{ran}\, \mathcal{J} = \mathrm{IS}$$

Of course in any implementation, we can avoid the rednndancy of having multiple copies of identical sub-trees by using pointers. In other words, in practice we represent a flowchart as a directed acyclic grapb, but mathematically we view the flowchart as a tree. Let F be our flowchart.

$$F : \mathrm{IT} \tag{7.10}$$

The corresponding concrete program is given hy

$$\psi_F \mathrel{\widehat{=}} (\mathcal{J} \fatsemi \mathcal{F} \fatsemi \mathcal{D} \fatsemi \mathcal{C})(F) \tag{7.11}$$

7.3 Implementation

As a notational convenience, we represent specifications as Boolean expressions over atomic propositions as we did for the rule-based system. Consider the compntational complexity of determining the fixed-point given a set of observations. Notice that an inference tree is finite (Equation 7.8). Let n be the number of distinct nodes in the tree. Since inference steps are decreasing, at each iteration the number of nodes whose truth valne is established must increase until a fixed-point is reached. This means that a fixed-point must be reached in no more than n steps. In the worst case, each iteration requires evaluation of all nodes in the tree. Therefore, the complexity of computing the fixed-point is $O(n^2k)$, where k is the worst-case time to evaluate an expression.

Unfortunately, evaluation of arbitrary expressions is computationally hard. In order to be able to use the flowchart efficiently in order to carry out diagnosis, we need to restrict

the class of expression that is allowed to appear in any decision node s. A convenient class is that of μ-expression: expressions in which each variable appears at most once. A μ-expression s can be easily evaluated if the current information t about the case is monomial. (A *monomial* expression is a μ-expression involving only conjunction.) Each proposition concerning a variable in s is evaluated independently given t, and then the truth values are combined using standard 3-valued logic operators. Since any input to ψ_F is a partial case, t is initially a monomial expression as required. In order that t remains monomial at each iteration, we insist that expressions appearing at leaf nodes are also monomial: the conjunction of any two monomial expressions is also monomial.

Chapter 8

Compilation of Database

This chapter describes the process of collecting cases and compiling a database for training and test purposes. A total of 1270 cases of abdominal or low back pain of suspected gynaecological origin were collected retrospectively from hospital case-notes.

8.1 A Medical Application

We have seen in previous chapters how different design decisions lead to a variety of diagnostic programs, ranging from purely statistical (e.g. the nearest neighbours method) to knowledge-based (e.g. a categorical flowchart). An important central question which should be asked by anyone contemplating the design and implementation of a diagnostic program is 'can the extra complexity and effort involved in building knowledge-based systems be justified in terms of a measurable improvement in diagnostic accuracy?' We address this by comparing all our programs on a suitable set of test cases. The application we have chosen is the diagnosis of abdominal pain. This has been one of the most widely studied applications [Fen90, Dom90], and it is regarded as the natural test field for further research into diagnostic methods [Sut89a]. In particular, we have confined our attention to a specific subgroup of patients, those for whom the pain is suspected to be of gynaecological origin. This was chosen because it is an especially challenging diagnostic task for both computer and clinician [Gun91]. Also the scope for exploiting knowledge of causal mechanisms appears to be greater amongst this selected group of patients: multiple pathophysiological states are common (e.g. pregnancy and a complication of pregnancy, chronic pelvic inflammatory disease and ectopic pregnancy, coincidental ovarian cyst and endometriosis etc.). If incorporation of background knowledge does improve the diagnostic accuracy of a program, then this should be more apparent in this selected group of patients. Clearly, however, any results in this highly specific application will need to be confirmed in other application areas. We have therefore designed our experimental methods and our computer programs with the objective of reuse.

8.1.1 Admission Criteria

Since we are interested in comparing different programs rather than in comparing computer with clinician, retrospectively collected cases are adequate for our purpose since any im-

pairment of the quality of clinical data disadvantages all programs equally. We therefore collected our cases retrospectively from hospital case-notes. Our admission criteria were

1. The patient was seen as an emergency by the admitting gynaecological team at the Churchill Hospital, Oxford, during the 12-month period 1/1/1990 to 31/12/1990 inclusive.

2. The age of the patient when seen was 13 years to 49 years inclusive.

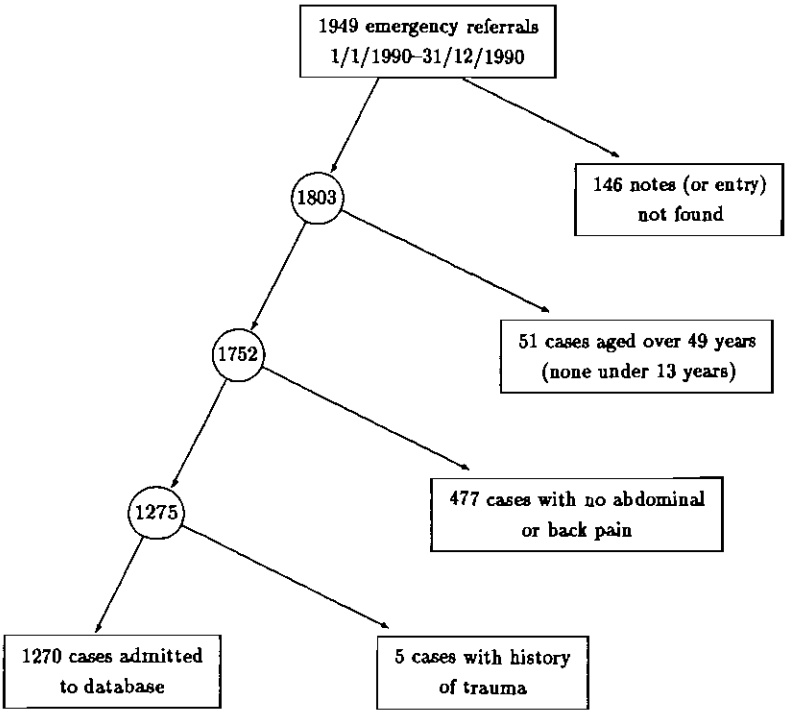
3. One of the patient's presenting complaints was abdominal or low back *pain* (not simply 'discomfort').

4. There was no history of trauma which could have explained the pain.

We chose to study a 12-month period in order to average out any conceivable seasonal variation in disease incidence, presentation or referral pattern. This also averages out the learning curves of Senior House Officers (who rotate every six months) in their ability to record symptoms and elicit signs. A total of 1949 emergency referrals were made in the 12-month period. The case-notes for 105 of these were unavailable, and it was uneconomical to continue to pursue them. In a further 41 cases, no relevant entry could be found in the case-notes, presumably because the case-notes had been unavailable at the time the patient was seen, and any temporary notes made had subsequently gone astray. This leaves 1803 cases in whom we found the clinical record. We included only patients of reproductive age because this reduces the range of diseases we needed to model. We note that the Leeds system did not perform well on children [Dic88] or in patients over the age of 50 years [Tel88], partly because conditions present differently in these groups. None of our 1803 cases were less than the lower age limit (13 years), but 51 were over the upper limit (49 years). A further 477 cases were eliminated because there was no recent history of abdominal or back pain (55 of these had perineal pain only, mostly due to Bartholin's cyst/abscess, another patient had chest pain only, and the rest had either no pain or just 'discomfort'). We included low back pain as an alternative to abdominal pain because pelvic pathology sometimes causes pain only in the back, and a decision to omit these patients would have seemed arbitrary. (About 3% of our patients had back pain but no abdominal pain.) In our selected group of patients the incidence of back pain due to serious non-gynaecological cause (e.g. vertebral disc protrusion) is negligible. Inclusion of back pain as a possible presentation therefore does not significantly complicate the model. We chose, however, to omit cases with a history of trauma because this small subset (5 cases) introduces many additional disorders. This leaves a total of 1270 cases which comprise our database. Figure 8.1 summarizes the above rejection frequencies.

Notice that the admission criteria do not exclude repeated presentations of the same patient during the study period. On average each patient occurs 1.3 times in the database. The case-notes of potentially admissible patients were retrieved in random order. However, in order to avoid requesting the same notes twice, data concerning any other admissible presentations of the same patient during the study period were recorded at the same time. This means that repeat presentations of the same patient tend to be near to one another in the database.

Figure 8.1: Rejection frequencies during data collection.



8.2 Final Diagnoses

The diagnostic task is to determine the fundamental cause of the patient's pain, which we refer to as the 'final diagnosis'. This is recorded for all cases as a single variable (corresponding to a in Equation 3.9). We drew up a list of 19 possible final diagnoses, chosen in order to separate patients into a relatively small number of major treatment and/or prognostic categories, while at the same time including all the common conditions. The decision as to which diagnoses to include was based both on those of the first 202 cases we collected, and on our expectations of future conditions we would encounter during data-collection. However, once we started to collect more cases we made no further modifications to the list. In order to determine the final diagnoses as reliably as possible we used the criteria which we enumerate below. In some cases, doubt remained about the true diagnosis. Rather than exclude such cases from the database, we qualified the diagnosis of every case with a label 'definite' or 'presumed'. If the case matched one of the following definitions, then the patient was assigned the corresponding diagnosis as a definite diagnosis. If the case matched none of the definitions precisely, then the patient was assigned the diagnosis whose definition was the closest match, as a presumed diagnosis. Thus no patient was excluded from the database because of uncertainty about the true diagnosis.

- **Non-Specific Pain** - Either the pain settles within 24 hours of admission without antibiotic therapy, or laparoscopy (or laparotomy) is performed. The patient is discharged without a definite cause for the pain being found, or with a vague and unreliable diagnostic label such as 'irritable bowel syndrome', 'Mittelschmerz' or 'dysmenorrhoea'. (The final diagnosis is presumed to be 'non-specific pain' if the patient takes her own discharge without a definite cause for the pain having been found, or if the patient is given antibiotics, but the clinicians' final diagnostic conclusion appears to be 'non-specific pain'.)

- **Threatened Abortion** - The patient is in the first 28/40 of pregnancy, and PV bleeding occurs at about the time of presentation, an ultrasound scan performed during the subsequent 14 days shows a viable intrauterine foetus, and the patient is discharged from hospital without the abortion becoming inevitable. (If an ultrasound scan shows a viable foetus and an intrauterine haemorrhage in the absence of PV bleeding, then the diagnosis of threatened abortion is presumed rather than definite.)

- **Abortion** - The patient is in the first 28/40 of pregnancy, and one of the following applies:

  - *(Missed)* - Evidence of foetal death *in utero* is found on ultrasound scanning. No other cause for the pain can be found.
  - *(Inevitable/Incomplete)* - There is no evidence that the abortion was missed. Evacuation of the uterus is undertaken during the current admission (even if the abortion appeared only to be threatened at presentation), and either products of conception are clearly identified at operation, or histopathological examination of curettage specimens confirms the presence of chorionic tissue.
  - *(Complete Abortion)* - The patient is found to have a spontaneously empty uterus following a confirmed pregnancy.

- **Retained Products** - Following a previous ERPC or termination of pregnancy or delivery, evacuation of the uterus is undertaken, and either products of conception are clearly identified at operation, or histopathological examination of curettage specimens confirms the presence of chorionic tissue.
- **Hydatidiform Mole** - A molar pregnancy is evacuated and confirmed histologically.
- **Ectopic Pregnancy** - An extrauterine pregnancy is removed surgically, and confirmed histologically.
- **Pelvic Inflammatory Disease** - At least one of the following criteria are satisfied in the presence of appropriate pelvic signs:
  1. The diagnosis is made on laparoscopy (or at laparotomy).
  2. The patient has either a pyrexia of at least 38°C or a white cell count of at least 15 per nanolitre, and responds rapidly (within 48 hours) to an appropriate antibiotic.
  3. Gonococcus is isolated from a high vaginal swab.
  4. Chlamydia antigen is found.

  In the case that pelvic inflammatory disease is secondary to retained products, the latter takes precedence as the final diagnosis.
- **Ovarian Cyst** - At least one of the following criteria is satisfied in the absence of any other explanation for the pain (and the clinicians appear to attribute the pain to the cyst):
  1. An ovarian cyst (on the same side as the pain if the pain is lateral) is found on ultrasound examination.
  2. An uncomplicated (i.e. neither torted, nor haemorrhagic nor ruptured) ovarian cyst (on the same side as the pain if the pain is lateral) is found at laparoscopy (or laparotomy).
- **Cystic Accident** - At least one of the following complications of an ovarian cyst is evident:
  - *(Adnexal Torsion)* - The diagnosis is confirmed at laparotomy.
  - *(Ruptured Cyst)* - The diagnosis is confirmed on laparoscopy or at laparotomy, or the patient has appropriate symptoms and signs, and free fluid in the Pouch of Douglas is found on ultrasound scanning, in the absence of other explanation for the pain.
  - *(Haemorrhage)* - The diagnosis is confirmed on laparoscopy or at laparotomy.
- **Pelvic Haematoma** - A pelvic haematoma is evacuated surgically.
- **Fibroids** - Fibroids are detected on ultrasound examination. Pain is associated with bleeding, or else a specific complication (e.g. red degeneration or torsion) is found on histological examination.
- **Hyperstimulation** - Bilateral ovarian enlargement is evident following therapeutic ovarian stimulation within the last month. No other cause can be found for the pain.

- **Urinary Tract Infection** - At least one of the following criteria is satisfied:
  1. The patient has dysuria or frequency, lower abdominal or loin tenderness, and at least moderate ('++') pyuria with no more than a minimal number ('+') of squames on urine microscopy.
  2. A pathogen is isolated from a mid-stream specimen of urine, no other cause for the pain can be found, and the patient responds rapidly to an appropriate antibiotic.
- **Endometriosis** - The diagnosis is confirmed at laparoscopy (or laparotomy).
- **Ureteric Colic** - A ureteric stone is passed spontaneously or confirmed radiologically.
- **Acute Appendicitis** - Histopathological examination of the appendix confirms appendicitis.
- **Hyperemesis Gravidarum** - Repeated vomiting during pregnancy is one of the presenting complaints, and the abdominal pain is aggravated by movement. No other cause for the vomiting or for the pain can be found.
- **Abdominal Wall Haematoma** - A haematoma is identified in the abdominal wall on ultrasound scanning, or is evacuated surgically.
- **Other** - Some definitive diagnosis other than those above is appropriate.

Table 8.1 shows the frequency distribution of the final diagnosis variable. The commonest diagnosis is 'abortion' (miscarriage): there are 468 cases, of which 372 are definite diagnoses. By contrast, ureteric colic and abdominal wall haematoma seem rare amongst this selected population; there is but a single example of each. The majority of patients have definite diagnoses, but the diagnosis that is most frequently presumed is that of 'pelvic inflammatory disease' (PID). The diagnostic criteria we have adopted for this are strong, and in practice the diagnosis tends to be made on clinical grounds alone. The patient then responds rapidly to antibiotic therapy, thus rendering laparoscopy unnecessary. Few patients have the necessary degree of pyrexia on presentation, and gonococcus and chlamydia are rarely confirmed as the causative organisms. Hyperemesis gravidarum is another condition which seldom satisfied our adopted diagnostic criteria since aggravation of pain by movement is usually not reported. In practice, though, there is usually no doubt about the diagnosis, and perhaps our diagnostic criteria could be relaxed.

8.3 Recorded Information

The data in the case-notes were transcribed onto a data-collection form using a formal protocol for interpreting the handwritten entries, and then transferred to computer. This allowed automatic range checking and scanning for other inconsistencies. Each case description was printed out and checked by hand against the information on the data collection form, so that any remaining errors could be corrected. The sources of data included clerking notes (even those made by medical students), referral letters, nursing records and discharge summaries. The objective was to record all information that was available (or *could* have been available had say an investigation been carried out in time) to the admitting team at

Table 8.1: Frequeucy distribution of final diagnosis.

Code Letter	Diagnosis	Definite	Presumed	Total
A	Non-specific pain	209	44	253
B	Threatened abortion	66	29	95
C	Abortion	372	96	468
D	Retained products	32	16	48
E	Hydatidiform mole	4	0	4
F	Ectopic pregnancy	69	3	72
G	Pelvic inflammatory disease	52	97	149
H	Ovarian cyst	19	20	39
I	Cystic accident	27	27	54
J	Pelvic haematoma	1	4	5
K	Fibroids	5	3	8
L	Hyperstimulation	0	3	3
M	Uriuary tract infection	4	8	12
N	Endometriosis	12	8	20
O	Ureteric colic	0	1	1
P	Appendicitis	2	0	2
Q	Hyperemesis gravidarum	4	10	14
R	Abdominal wall haematoma	1	0	1
S	Other	16	6	22
	(total)	895	375	1270

the time of the patient's presentation. We recorded values for up to 169 variables describing relevant symptoms and other historical details (age, previous operations etc.), physical signs and the results of investigations (blood and urine tests, and ultrasound scan). We refer to all of these as *symptom variables*. On average, values were recorded for 62% of the symptom variables in any given case. The protocol we used for recording each variable is summarized in the appendix (Appendix A), and the possible values for each symptom variable are shown in Tables A.1, A.2, and A.3.

A note was also kept of other information such as the source of referral, and the clinicians' initial diagnoses. However, these were not included as symptom variables to assist computer diagnosis because they reflect decisions by others as to the cause of the symptoms. For example, a patient tends to be referred from the ultrasound clinic because a scan has shown a missed abortion or an ectopic. The initial diagnosis was recorded as the provisional diagnosis made at presentation. If a list of possible diagnoses was drawn up, then only the first in the list was recorded. Question marks qualifying diagnoses were ignored. Where necessary, the diagnosis was translated to one relevant to the pain rather than to other symptoms or signs.

We also recorded values for a further 53 variables which represent various pathophysiological states and refinements of the final diagnosis. We refer to these as *additional variables*. Any operative findings were particularly useful in determining the underlying causal mechanism of the patient's symptoms and signs, but in many cases a subjective decision had to be made. In all but a few exceptional cases, the values of all the additional variables could be decided and recorded, so missing values are rare except for variables which are conditional on others. Additional variables are not used for test purposes (their values like that of the final diagnosis are concealed in any test case). However, the values of the additional variables in training cases are made available to any programs which can exploit them. The protocol we used for recording each additional variable is also summarized in the appendix (Appendix A), and the possible values for each additional variable are shown in Table A.4.

8.4 Criticism of Our Choice of Variables

During the process of data-collection, we kept a note of any perceived inadequacies of our choice of variables to record. The following is an itemized list of these comments. Although the list is long, most of the points are minor. Perhaps the most significant improvement would be to record the severity of vomiting as suggested.

1. It should be possible to describe chronic pain separately from acute pain. For example, the patient may have had low-back pain for a month, but, in the last few days experienced central abdominal pain moving to the RIF. At present a subjective decision has to be made as to whether the low-back pain is part of the present complaint, in which case 'back' rather than 'RLQ' is recorded as the initial site.

2. Pain sometimes radiates to the right hypochondrium, or chest. No variables are available specifically for this: the closest is 'epigastrium'.

3. The sensitivity of the pain to movement should be more clearly defined: at present no distinction is made between pain that is made worse by any attempt at movement, and pain that is simply aggravated by walking.

4. It is sometimes hard to distinguish mild pain from 'discomfort'.

5. A way should be found of describing 'cramping, intermittent pain': at present these two adjectives are alternatives.

6. Both the type of the pain at onset, and the type of the pain at presentation should be recorded in case it has changed.

7. Any changing menstrual pattern should he recorded.

8. A distinction should be made between 'long' LMP and 'heavy' LMP, and 'short' LMP and 'light' LMP.

9. A variable should record whether abnormal PV bleeding preceded, started at the same time as, or followed the pain. This can help distinguish ectopic pregnancy from abortion.

10. A distinction should he made between fresh blood loss and brownish loss; this is useful in the diagnosis of missed abortion.

11. Any 'sensation of pregnancy' should be recorded: this is usually present even in the case of ectopic pregnancy, and it often disappears in the case of a missed abortion.

12. A distinction should be made between primary and secondary dysmenorrhoea.

13. A distinction should be made between actual fainting and the sensation of faintness.

14. The *severity* of vomiting should be recorded; this is an essential and obvious clue in the diagnosis of hyperemesis gravidarum.

15. A distinction should be made between an episode of some symptom (e.g. vomiting) a week or so ago, and the presence of the symptom for the last week or so. At present only the time since onset is recorded.

16. Previous cervical smear results should be recorded.

17. If the patient is pregnant, and the normal method of contraception is IUCD, then it should be recorded whether the IUCD has been removed.

18. Any past history of hyperemesis in the present or previous pregnancies should be recorded.

19. It should be made clear whether the patient has delivered (or had TOP, ERPC or complete abortion) *since the LMP.*

20. Previous myomectomy should be recorded.

21. There should just be a single variable 'uterine instrumentation', and this should record *all* instrumentation, whether dnring laparoscopy or not.

22. A history of previous termination of pregnancy should be recorded in a single variable, and not duplicated in the variables 'terminations' and 'previous termination' as it is at present.

23. Beta blockers should be included in the drug history; they may mask tachycardia.

24. The respiratory rate is usually available, and should be recorded.

25. A distinction should be made between facial pallor and clinical anaemia; facial pallor is a sign both of a sympathetic response and of anaemia.

26. An extra value 'in pain' should be included for the 'mood' variable; at present this is recorded simply as 'other'.

27. A distinction should be made between a suspected abdominal mass and one that is definite, in the same way as for PV examination.

28. Central PV tenderness should be divided into 'tender uterus' and 'tender POD'.

29. PV masses are sometimes felt centrally; at present there is no way of recording these.

30. The severity of cervical excitation should be recorded.

31. The results of both high and low sensitivity pregnancy tests should be recorded if known; a negative low sensitivity test in the presence of an established pregnancy conveys useful information.

32. The pH of the urine should be recorded.

33. The results of electrolytes, liver function tests and amylase blood tests should be recorded.

34. The results of ultrasound examination of the gallbladder and kidneys should be recorded.

35. The results of IVP investigation should be recorded.

36. An additional variable should record the presumed nature of the uterine contents.

Chapter 9

Construction and Validation of Knowledge Bases

This chapter describes the construction and validation of the exemplar model, Bayesian networks, flowchart, and rule-based systems.

9.1 Causal Models

In constructing the knowledge bases we utilized information from various sources: standard textbooks (e.g. [Pau82, Cha84, Whi86]), journal articles, personal experience and discussion with medical colleagues. We also made a careful study of the first 202 cases we collected, seeking a full causal explanation for each one in order to identify the most important causal mechanisms. (Consequently these 202 cases could not be used subsequently for test purposes.) In some instances, as we describe, machine assistance could be provided in the task of learning from these cases. We also used all 1270 cases to criticize various knowledge bases by means of a χ^2 test of goodness of fit. (This criticism is retrospective: no changes to the knowledge bases were made in the light of the χ^2 tests.) We describe first the development of the causal knowledge bases, and afterwards the development of the inferential knowledge bases.

9.1.1 Exemplar Model

The simplest knowledge base of all is that for the exemplar method. Construction of this entailed drawing up 19 typical symptom profiles, one for each disease. To assist in this process, we implemented a program to display the frequency distribution for each symptom variable given each disease amongst the 202 cases. For example, there were 16 cases of ectopic pregnancy, and the frequency distribution for the variable 'type of pain' is shown in Table 9.1. Note that the variable was recorded in only 10 of the 16 cases. In this instance the data confirmed our expectation that the typical patient with an ectopic pregnancy describes 'cramping' abdominal pain. Often though, the data conflicted with our expectations. If numbers were small, we tended to adhere to our expected value. For example, although in only one case of ectopic pregnancy did the urine contain no pus cells whereas they were found in three other cases (Table 9.2), we considered it more typical for the urine to he free

of pus cells, and we took 'none' as the typical value. When numbers were larger, we tended to choose the mode of the frequency distribution even if this was counter to intuition. For example, during the first 12 weeks of an ectopic pregnancy, the uterus grows to nearly the same size as it would in an intrauterine pregnancy [Cun89]. One would expect, therefore, that the uterus would be noted to be enlarged in cases of ectopic pregnancy. In practice this appears not to happen (Table 9.3). We therefore took the typical value for the variable 'uterus enlarged' to be 'false' in cases of ectopic pregnancy.

Table 9.1: Frequency distribution of the variable 'type of pain' amongst the 10 cases of ectopic pregnancy in which this variable was recorded.

Value	Frequency
intermittent	2
steady	1
colicky	2
cramping	4
fluctuating	1
other	0
(total)	10

Table 9.2: Frequency distribution of the variable 'urine microscopy pus cells' amongst the 4 cases of ectopic pregnancy in which this variable was recorded.

Value	Frequency
none	1
minimal	2
moderate	1
(total)	4

Table 9.3: Frequency distribution of the variable 'uterus enlarged' amongst the 13 cases of ectopic pregnancy in which this variable was recorded.

Value	Freqnency
false	9
true	4
(total)	13

One of the difficulties we encountered in trying to construct templates is that many of the disease categories are heterogenous. Several conditions can present with a right or left

mirror image (ectopic pregnancy, pelvic inflammatory disease, ovarian cyst, cystic accident, urinary tract infection, endometriosis, ureteric colic). A somewhat arbitrary choice has therefore to be made as to whether the more typical presentation is with left or right-sided symptoms and signs. Other conditions (e.g. abortion) are a heterogenous mixture of more specific types (i.e. inevitable abortion, incomplete abortion, complete abortion, and missed abortion). The choice as to which is the more typical is also rather arbitrary. This suggests that the exemplar model should be refined to the point at which subcategories can no longer be usefully distinguished, and a separate template constructed for each such subcategory. Since the present 19 templates have already required careful consideration of 3211 disease-symptom pairs, we have not yet refined the exemplar model further.

Model Criticism

Retrospectively we are able to criticize the exemplar model in the light of our experience of 1270 cases. A simple means to do this is to use all 1270 cases for training purposes to calculate the necessary parameters (the various $p(\Delta = d \mid \{\})$ and $p(v : u \rightsquigarrow u')$) and then perform a χ^2 goodness of fit test of the observed distribution of each symptom variable to that predicted. This measures the degree to which the exemplar model is able to adapt to the given training sample. (We measure the model's predictive ability on unseen cases in the next chapter.)

Of the total of 3211 disease-symptom pairs, 256 failed the χ^2 test at the 1% threshold. (We would expect only about 32 to fail by chance.) The worst pair was the variable 'clinically dehydrated' given the diagnosis of hyperemesis gravidarum: see Table 9.4. The χ^2 statistic is 198.00 while the 1% threshold is only 6.63 (one degree of freedom). Clearly the origiual decision was correct that patients with hyperemesis gravidarum are not typically dehydrated: only 5 of the 14 cases amongst the 1270 were found to be dehydrated. However, dehydration is nevertheless characteristic of this condition, and rare in the other diagnostic classes. Had the template for hyperemesis gravidarum recorded the typical value of 'clinically dehydrated' as 'true' rather than as 'false' then an almost perfect fit would have been obtaiued ($\chi^2 = 0.03$). This suggests that in constructing the templates it would have been better to have chosen the most 'characteristic' values (i.e. those which are most suggestive of the given disease) rather than the most typical values. This is an area worthy of future investigation.

Table 9.4: Comparison of the actual frequency distribution of the variable 'clinically dehydrated' with that expected according to the exemplar model, amongst the 14 cases of hyperemesis gravidarum in which this variable was recorded. The χ^2 statistic for this variable is 198.00, while the 1% significance threshold is 6.63.

Value	Predicted Probability	Expected Frequency	Actual Frequency
false	0.9913	13.88	9
true	0.0087	0.12	5
(total)	1.0000	14.00	14

9.1.2 Bayesian Networks

We set out to construct a range of Bayesian networks of varying complexity in order to gauge the contribution of background knowledge. We already had a trivial network, 'independence Bayes' (parents relation P_I). We therefore implemented a large network (parents relation P_L), and then by simplification we derived a smaller network (parents relation P_S) from it. This gave us a set of three networks, ranging from the trivial to the complicated. We describe first the construction of the large network.

The essential difficulty we encountered was that the final diagnosis variable Δ has 19 possible values, yet all symptoms and signs depend on the final diagnosis. The symptom variable 'final site of pain', for example, has 14 values and so there are 247 $(= 19 \times (14 - 1))$ parameters in its conditional probability table. If we introduce into the network additional variables such as 'uterine contractions' (5 values) and 'acute left pyelonephritis' (5 values) then since many symptoms and signs depend on these too, the associated conditional probabilities become far too numerous either to estimate or store.

We sought to offset these difficulties by two coding tricks. Firstly, we often made causes the children of effects rather than the parents. For example, in our network 'past history of PID' is a child of 'final diagnosis' rather than a parent. Had all ten binary 'past history' variables been parents of 'final diagnosis' then the conditional probability table for the latter would have had 18432 $(= 2^{10} \times (19 - 1))$ parameters! Instead we have 10 tables, each with just 19 $(= 19 \times (2 - 1))$ parameters. Secondly, we created new additional variables 'pathological process' (22 values) and 'anatomical process' (32 values) which represent refinements of 'final diagnosis' (19 values), thereby avoiding the multiplication of table size that occurs if additional variables share the role of parent with 'final diagnosis'. The values of these new variables were easily determined from the other additional variables that we had collected.

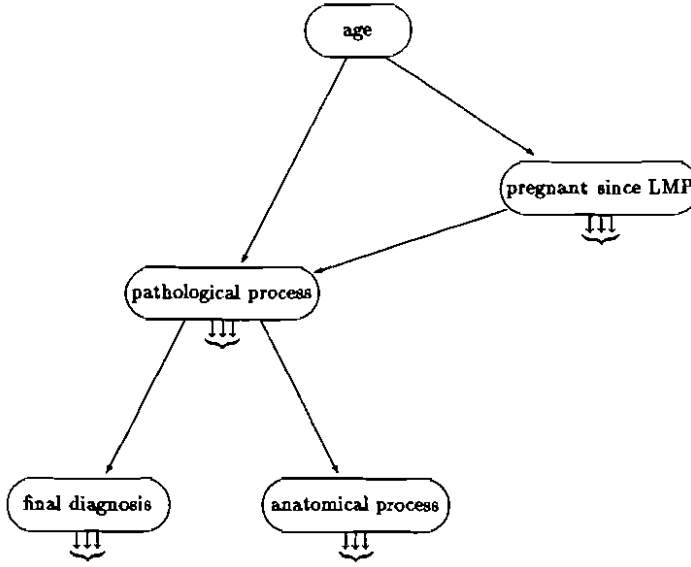
The variable 'pathological process' conveys more information than 'final diagnosis' about the stage of abortion and type of urinary tract infection: 'abortion' is replaced by three new values, 'inevitable abortion', 'incomplete abortion' and 'complete abortion'; 'urinary tract infection' is replaced by 'acute pyelonephritis' and 'acute cystitis'. We felt that it was important to refine the heterogenous conditions 'abortion' and 'urinary tract infection' because the former is so common and the latter presents in essentially two very different ways. The variable 'pathological process' is thus an alternative to 'final diagnosis' as a parent for variables in which the stage of abortion or type of urinary tract infection is significant (e.g. 'progress of pain'). Similarly, the variable 'anatomical process' refines conditions which have left/right-sided varieties. See Table 9.5 for details.

At the root of the Bayesian network we placed the variable 'age'. This is clearly a cause rather than an effect of any other variable. Many of the possible diseases that we are modelling are complications of pregnancy. We therefore introduced a new binary variable 'pregnant since LMP' to indicate whether the patient became pregnant since her last menstrual period. We then placed the 'pathological process' as a child of these two variables, and 'final diagnosis' and 'anatomical process' as children of 'pathological process' (Figure 9.1). We could have put 'anatomical process' in the place of 'pathological process', but the former has 32 values compared with the latter's 22, and so the size of the associated conditional probability table would have been 248 $(= 4 \times 2 \times (32 - 1))$ rather than 168 $(= 4 \times 2 \times (22 - 1))$. This would have made less efficient use of the available training data.

Table 9.5: Refinement of 'final diagnosis' into 'pathological process' and 'anatomical process' for use in the Bayesian networks.

final diagnosis	pathological process	anatomical process
non-specific pain	non-specific pain	non-specific pain
threatened abortion	threatened abortion	threatened abortion
abortion	inevitable abortion incomplete abortion complete abortion	inevitable abortion incomplete abortion complete abortion
retained products	retained products	retained products
hydatidiform mole	hydatidiform mole	hydatidiform mole
ectopic pregnancy	ectopic pregnancy	left ectopic pregnancy right ectopic pregnancy
pelvic inflammatory disease	pelvic inflammatory disease	left PID right PID bilateral PID
ovarian cyst	ovarian cyst	left symptomatic ovarian cyst right symptomatic ovarian cyst bilateral symptomatic ovarian cysts
cystic accident	cystic accident	left cystic accident right cystic accident
pelvic haematoma	pelvic haematoma	pelvic haematoma
fibroids	fibroids	fibroids
hyperstimulation	hyperstimulation	hyperstimulation
urinary tract infection	acute pyelonephritis acute cystitis	acute left pyelonephritis acute right pyelonephritis acute cystitis
endometriosis	endometriosis	left symptomatic endometriosis right symptomatic endometriosis bilateral symptomatic endometriosis
ureteric colic	ureteric colic	left ureteric colic right ureteric colic
acute appendicitis	acute appendicitis	acute appendicitis
hyperemesis gravidarum	hyperemesis gravidarum	hyperemesis gravidarum
abdominal wall haematoma	abdominal wall haematoma	abdominal wall haematoma
other	other	other

Figure 9.1: Nodes at the root of the large Bayesian network.



One of the commonest and most important symptoms of patients presenting to the gynaecologist with abdominal (or back) pain is associated PV bleeding and menstrual disturbance. Naturally patients tend to confuse abnormal bleeding with a normal menstrual period, and a significant part of the diagnostic task is to interpret when the actual LMP really occurred. We were able to model this quite simply. We introduced two kinds of node: additional variables to record the time of the actual LMP and any abnormal bleeding, and symptom variables to record the description given by the patient. A third new additional variable, 'reported LMP', records the temporal relationship ('earlier', 'same' or 'later') between the actual and reported time of the LMP. The LMP may be falsely identified and reported later than the actual LMP if abnormal uterine bleeding has occurred since the actual LMP, or if an implantation haemorrhage has occurred. The latter of course occurs only if the patient became pregnant since the LMP, and the site of the pregnancy was uterine. The description of the LMP and of any bleeding since depends on the relative timing of the actual and reported LMP. For example, if an implantation haemorrhage has occurred, and if the reported LMP was later than the actual, then it is likely that the implantation haemorrhage was mistaken for the LMP. This means that the LMP will tend to be described as occurring earlier than expected and of lighter flow than normal. The part of the network that models these interdependencies is shown in Figure 9.2. Making similar reference to 'reported LMP', other variables not shown model the relationship between the actual and reported severity and progress of any bleeding and the passage of product of conception. A similar problem arises in interpreting the presence of red cells in the urine. Usually they are due to contamination from PV bleeding, but occasionally they may actually signify haematuria. Figure 9.3 shows how this was modelled. A similar approach was taken in handling the presence of pus cells in the urine due to contamination.

For the most part, the other symptom variables have a single parent: either 'final diagnosis', or 'pathological process', or 'anatomical process', depending on whether the extra information conveyed is relevant. Some variables which also represent physiological effects of pregnancy (e.g. 'constipation', 'frequency' and 'discharge') also have 'pregnant since LMP' as a parent.

The Small Bayesian Network

We derived the small Bayesian network from the large. We dispensed with all additional variables except 'pathological process' and 'anatomical process'. All symptom variables with a few exceptions thus had a single parent. The exceptions are shown in Figure 9.4. These represent obvious interactions between symptom variables. Two variables, 'urine microscopy squames' and 'ultrasound type', have no parents. This is because there is no obvious causal dependence of either on the underlying disease process. Squames in the urine merely indicate that the specimen was contaminated, this makes it more likely that pus cells will be found too. The decision as to whether to perform an abdominal or vaginal ultrasound is a clinical one; we therefore preferred not to take this into account as direct evidence for any particular diagnosis. However, the pelvic structures are more easily examined by vaginal ultrasound, so the type of ultrasound affects the likelihood of detection of abnormalities. The type of ultrasound is therefore a parent of several other ultrasound variables as shown in Figure 9.4. The number of uterine pregnancies detected (if any) is dependent on any recent fertility therapy and, as far as we know, nothing else. Recent fertility drugs are much more likely to have been administered if the patient gives a history of infertility. We

Figure 9.2: Part of the large Bayesian network which models the influence of an implantation haemorrhage and/or abnormal uterine bleeding on the reported dates of the LMP. (The symbols $\overset{\frown}{\downarrow\downarrow\downarrow}$ and $\underset{\smile}{\downarrow\downarrow\downarrow}$ indicate the presence of one or more other arcs entering and leaving the node, respectively.)

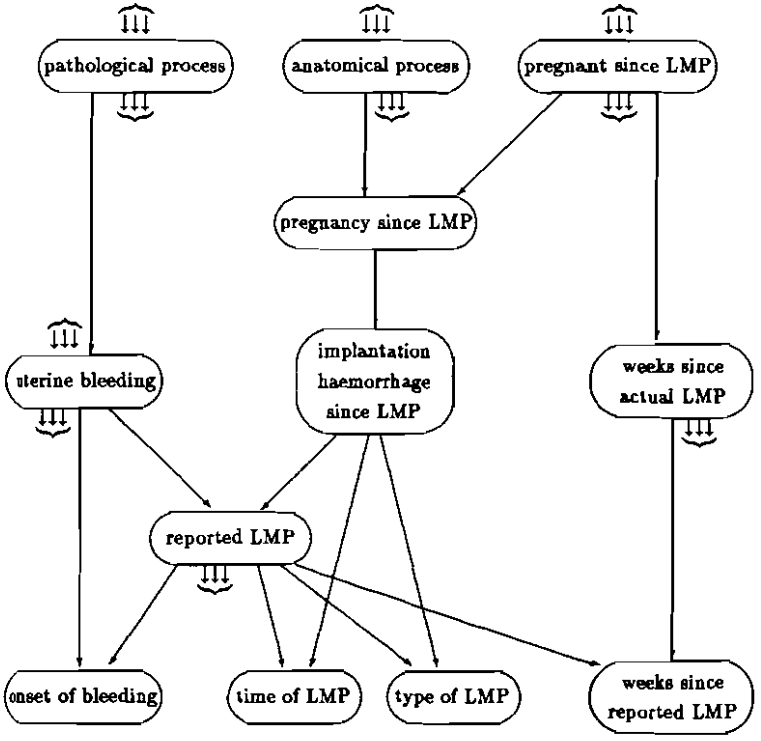
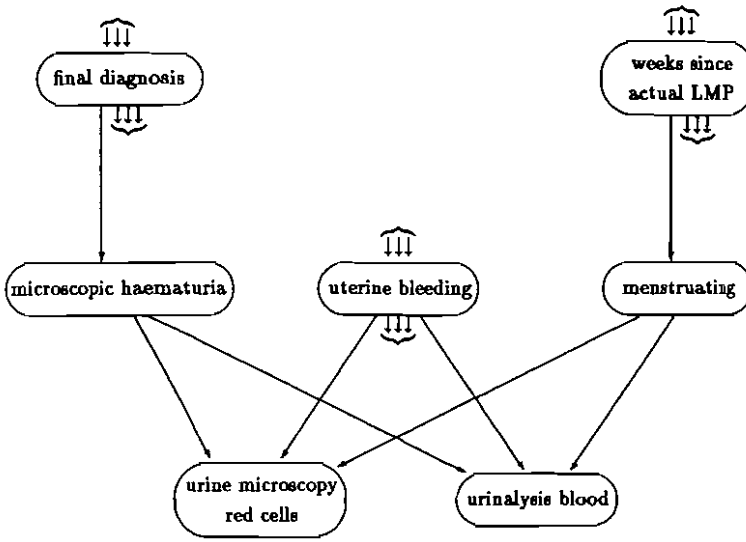


Figure 9.3: Part of the large Bayesian network which models the influence of haematuria and PV bleeding on the finding of red cells in the urine either on microscopy or on stick testing.



therefore included 'infertility' as a parent of 'recent fertility drugs'. Lastly, the reported time since onset of abnormal bleeding is clearly dependent on the reported time since the last menstrual period: by definition it must be less. We therefore included 'weeks since reported LMP' as a parent of 'onset of bleeding'.

The small Bayesian network has significantly fewer arcs than the large network, but is still more complex than the independence network. Table 9.6 shows a comparison of the numbers of nodes, arcs and parameters in each of the three Bayesian networks. The number of parameters was calculated according to

$$\text{Parameters} = \sum_{v:\text{Var}} \left((1 - \#\theta(\lfloor v \rfloor)) \prod_{v':\text{Var}|(v,v')\in P} \#\theta(\lfloor v' \rfloor) \right)$$

where P is the corresponding parents relation. Clearly the numbers of parameters of all three Bayesian networks are of a similar order of magnitude, and not excessive for the size of database available for training purposes.

Table 9.6: A comparison of the numbers of nodes, arcs and parameters of the three Bayesian networks: the independence network (P_I), the small network (P_S) and the large network (P_L).

Network	Nodes	Arcs	Parameters
P_I	170	169	7333
P_S	172	176	10086
P_L	185	225	11471

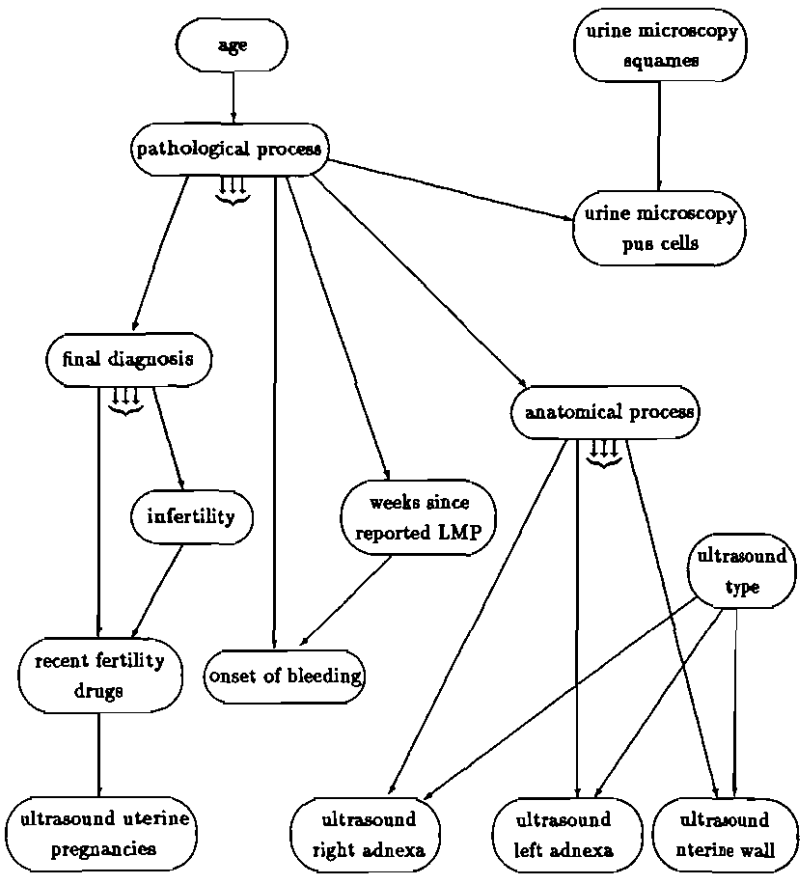
9.1.3 Causal Rule-Based System

The greater flexibility of the rule-based representation made it unnecessary to introduce special variables such as 'anatomical process', 'pregnant since LMP' and 'reported LMP' in order to restrain growth in the number of parameters. For example, instead of the new binary variable 'pregnant since LMP' we were able to write an equivalent atomic proposition in terms of the existing additional variable 'pregnancy since LMP'.

```
pregnant_since_LMP = true
    ≡
pregnancy_since_LMP  in { uterine,
                          left_tubal,
                          right_tubal,
                          left_ovarian,
                          right_ovarian,
                          hydatidiform_mole }
```

We created only one new variable 'recent TOP or ERPC' which records whether or not the patient has had a recent termination of pregnancy, or evacuation of retained products of

Figure 9.4: Part of the small Bayesian network omitting only symptom variables with single parents.



conception. It was necessary to introduce this variable to avoid repetition: it records an important fact about the case, which is used repeatedly in subsequent families of rules, yet no equivalent variable was available amongst the existing set of additional variables. The total number of variables (symptom variables, additional variables and the disease variable) is thus 224.

We constructed the knowledge base by first drawing up a causal sequence Q_C of 478 atomic propositions over these variables. The sequence begins with propositions concerning age and historical details such as previous surgery. It then covers pathophysiological states and finally symptoms, signs and results of investigations. We have therefore arranged causative entities early in the sequence, and their observable effects later in the sequence. The sequence is thus a true causal ordering; since the rule-based representation avoids combinatorial explosion in the number of parameters, the sort of coding trick employed in the Bayesian network was not needed. The sequence is, however, not complete: for example, we preferred in many cases not to refine the times of events beyond distinguishing whether they are recent (less than a month) or distant (more than a month). Thus the variable 'time since appendicectomy' has five possible values: 'hours', 'days', 'weeks', 'months' and 'years'. However, sequence Q_C includes only the following proposition about this variable.

$$\text{time\_since\_appendicectomy} \ \ in \ \{ \text{ hours, days, weeks } \}$$

Therefore no preference is expressed amongst these values, or between 'months' and 'years'. By default, probability is distributed uniformly over each such set (Equation 6.17).

Our next task was to write a family of rules for each of the atomic propositions in Q_C. In formulating the rules we limited the number in each family in accordance with the anticipated amount of training data available to derive the certainty factors. First we included any such categorical rules as were necessary to express logical dependence of the proposition on those anterior to it in Q_C, then afterwards we included uncertain (non-categorical) rules. We arranged the latter as far as possible to represent separate pieces of evidence. The average number of rules associated with each proposition is between four and five: there are a total of 2143 rules in the knowledge base, of which 571 are categorical (506 logical preclusions and 65 logical implications).

For example, shown below is the family of five rules we formulated to determine the probability of facial pallor. The certainty factors shown are those derived by iterative optimization (except for Rule 9.1.1 which is categorical) using the entire database of 1270 cases as a training set.

Rule 9.1.1

$$\text{colour} \ \ in \ \{ \text{ flushed } \} \Rightarrow^{0} \text{colour} \ \ in \ \{ \text{ pale } \}$$

This first rule is categorical. It dictates that if the patient has a flushed face then pallor cannot simultaneously be present. This rule is required in order to ensure external consistency (Equation 6.14).

Rule 9.1.2

$$\text{true} \ \Rightarrow^{0.043} \text{colour} \ \ in \ \{ \text{ pale } \}$$

This reflects the relatively low prevalence of facial pallor amongst our group of patients.

Rule 9.1.3

mean_bp in { less_than_70mmHg } $\Rightarrow^{0.796}$ colour in { pale }

This identifies hypotension as a cause of facial pallor.

Rule 9.1.4

peritoneal_cavity in { moderate_haemoperitoneum, massive_haemoperitoneum }
or
(uterine_bleeding in { hours, days, weeks, months }
and
severity_of_uterine_bleeding in { heavy }
)
$\Rightarrow^{0.711}$ colour in { pale }

This identifies significant blood loss as a cause of pallor, whether the bleeding is internal or external.

Rule 9.1.5

mood in { anxious }
or
severity_of_pain in { severe }
$\Rightarrow^{0.886}$ colour in { pale }

This identifies increased sympathetic tone due to anxiety or pain as a contributory factor.

The rules represent different kinds of evidence, although the antecedents share a common causal pathway: increased sympathetic tone. Moreover, massive haemorrhage (Rule 9.1.4) causes hypovolaemia and hypotension (Rule 9.1.3). Fortunately, the logistic model does not require conditional independence. The antecedents may even share variables as the next example shows. This is permissible because the logistic formula is consistent with logical dependence (exclusion or implication) between its terms. Shown below is the family of six rules which determine whether the left adnexa appears abnormally enlarged on ultrasound examination. Three kinds of abnormality are represented in the simulation model; the adnexa may simply appear enlarged, or a solid mass may be detected, or a cyst may be seen. No distinction is made between these three possibilities in this family of rules, other families carry out that function.

Rule 9.1.6

previous_left_salpingectomy in { true }
and
previous_left_oophorectomy in { true }
$\Rightarrow^{0}$ ultrasound_left_adnexa in { enlarged, mass, cyst }

This first rule is categorical. If both the left Fallopian tube and the left ovary have been previously removed then obviously no left adnexal enlargement (of any kind) is possible.

Rule 9.1.7

true $\Rightarrow^{0.025}$ ultrasound_left_adnexa in { enlarged, mass, cyst }

This rule reflects the fact that the left adneza usually appears normal on ultrasound examination.

Rule 9.1.8

left_ectopic_pregnancy in {unruptured,
ruptured_into_mesosalpinx,
ruptured_into_peritoneal_cavity}
$\Rightarrow^{0.898}$ ultrasound_left_adnexa in { enlarged, mass, cyst }

The presence of a left ectopic pregnancy (irrespective of whether or not it is ruptured) makes it much more likely that some form of enlargement of the left adneza will be detected.

Rule 9.1.9

left_ovarian_cyst in { asymptomatic, symptomatic, haemorrhagic, ruptured, torted }
$\Rightarrow^{0.991}$ ultrasound_left_adnexa in { enlarged, mass, cyst }

The presence of a left ovarian cyst makes it very much more likely that some form of enlargement of the left adneza will be detected. It is irrelevant whether the cyst is symptomatic or not, and whether it is complicated in some way.

Rule 9.1.10

left_hydrosalpinx in { true }
or
left_pyosalpinx in { true }
$\Rightarrow^{0.985}$ ultrasound_left_adnexa in { enlarged, mass, cyst }

Fluid (whether purulent or not) in the left Fallopian tube makes detectable adnexal enlargement of some kind very much more likely.

Rule 9.1.11

(left_ectopic_pregnancy in {unruptured,
ruptured_into_mesosalpinx,
ruptured_into_peritoneal_cavity}
or
left_ovarian_cyst in { asymptomatic, symptomatic, haemorrhagic, ruptured, torted }
or
left_hydrosalpinx in { true }
or
left_pyosalpinx in { true }
)
and
ultrasound_type in { vaginal }
$\Rightarrow^{0.706}$ ultrasound_left_adnexa in { enlarged, mass, cyst}

Adnexal enlargement due to any pathology is more likely to be detected if the ultrasound is performed vaginally rather than abdominally since fewer structures shield the pelvic organs from the transducer.

The certainty factors of the 1572 non-categorical rules were derived from the given training sample by iterative maximum likelihood estimation, using simple gradient descent with a gain of unity. With respect to each family of rules, only cases in which the truth value of the conclusion was known were used for training purposes. There were 1268 and 762 such training cases, respectively, for the two families of rules shown above. When evaluating the antecedent of a rule during training, each component proposition 'v in U' was assumed to be false (i.e. have value '0') if the value of variable v was unrecorded: the rationale for this was that significant diagnostic features would have been recorded had they been present. Figure 9.5 plots mean surprise as a function of training iteration for both the families of rules shown above. Clearly there is no significant reduction in surprise after about 500 iterations, and 1000 iterations would therefore appear to be adequate.

Figure 9.5: Graph of average surprise per training case as a function of training iteration for the two families of rule in the causal rule-based system with conclusions as shown.

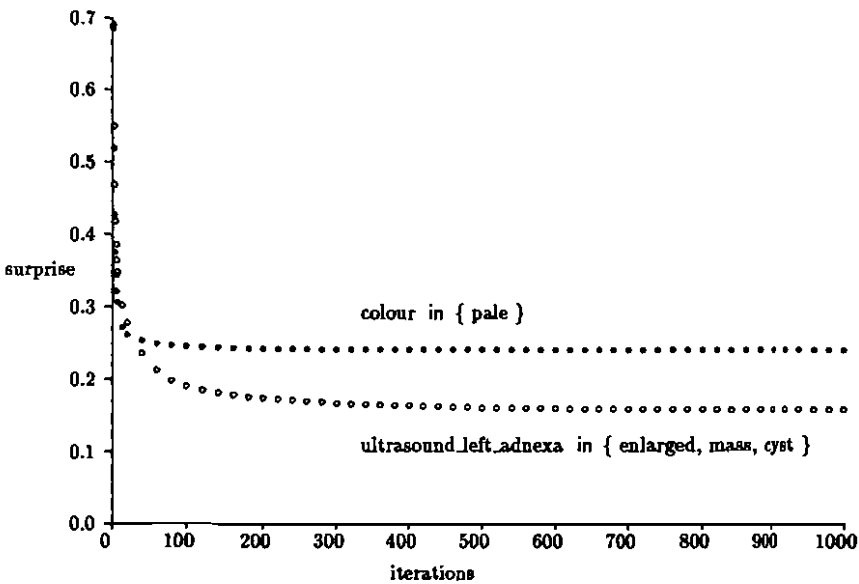


Table 9.7 shows the frequency distribution of all 1572 non-categorical certainty factors. The distribution is roughly uniform over the interval $[0, 1]$, showing that the full range of certainty factors is used. Tables 9.8 and 9.9 enumerate the conditional probabilities defined by the two families of rules shown above. In both cases, although there are roughly twice

as many data as there are degrees of freedom (rules with non-categorical certainty factors), a reasonable fit is obtained.

Table 9.7: Frequency distribution of the 1572 non-categorical certainty factors in the causal rule-based system derived from the entire database of 1270 cases.

Interval	Frequency
(0.0, 0.1]	207
(0.1, 0.2]	97
(0.2, 0.3]	91
(0.3, 0.4]	123
(0.4, 0.5)	152
[0.5, 0.6)	201
[0.6, 0.7)	191
[0.7, 0.8)	180
[0.8, 0.9)	162
[0.9, 1.0)	168
(total)	1572

9.1.4 Chi-Square

A more general test of the rule-based system's ability to fit the training data is to compare the specified and observed marginal distributions of each variable by means of the χ^2 test, just as we did for the exemplar model. (There was little point in carrying out this test for Bayesian networks, because the specified marginal distribution of any variable necessarily conforms to that observed.) We estimated the marginal distributions specified by the model for each of the 224 variables by generating 10^6 cases, and counting relative frequencies. Then we used the χ^2 test to falsify the hypothesis that the same database of 1270 real cases used to train the model is a random sample generated from the model. At the 1% significance level, 28 variables failed the χ^2 test. They are shown in Table 9.10.

Many variables fail the χ^2 test because of incompleteness of the sequence Q_C. These include variables relating to the time since operations, type of contraception, type of pregnancy since the LMP, raised progesterone level, and ovarian cysts. The default assumption of uniform distribution of probability within each equivalent set of values is inconsistent with the observations. For example, it was felt that a history of previous cervical surgery made little contribution to the diagnosis of abdominal pain, and so no proposition concerning this variable was included in Q_C. As a result, probability is distributed uniformly over its two values, yet fewer than 5% of patients have actually had previous cervical surgery: see Table 9.12. Similarly, no distinction was made in the model between an ovarian and a tubal ectopic pregnancy, because it was felt that the difference was not significant diagnostically, and therefore not worth modelling: see Table 9.14. It is therefore not surprising that these variables fail the χ^2 test. (In an earlier paper, we compared only the distributions over equivalence classes of values, so this effect was not observed [Tod93b]).

Table 9.8: Conditional probabilities of 'colour in { pale }' computed by the corresponding family of rules: Rules 9.1.1 to 9.1.5. (Note that Rule 9.1.2 always fires because its antecedent is 'true'.) Also shown are the relative frequencies with which the conclusion holds given each pattern over the training set of 1268 cases.

Pattern of Rule Firing					Computed	Relative	
9.1.1	9.1.2	9.1.3	9.1.4	9.1.5	Probability	Frequency	
0	1	0	0	0	0.043	33/ 748	= 0.044
0	1	0	0	1	0.090	21/ 250	= 0.084
0	1	0	1	0	0.100	19/ 179	= 0.106
0	1	0	1	1	0.196	9/ 48	= 0.188
0	1	1	0	0	0.150	1/ 13	= 0.077
0	1	1	0	1	0.278	3/ 4	= 0.750
0	1	1	1	0	0.302	2/ 9	= 0.222
0	1	1	1	1	0.486	1/ 2	= 0.500
1	1	0	0	0	0.000	0/ 10	= 0.000
1	1	0	0	1	0.000	0/ 4	= 0.000
1	1	0	1	0	0.000	0/ 0	
1	1	0	1	1	0.000	0/ 1	= 0.000
1	1	1	0	0	0.000	0/ 0	
1	1	1	0	1	0.000	0/ 0	
1	1	1	1	0	0.000	0/ 0	
1	1	1	1	1	0.000	0/ 0	
					(total)	89/1268	= 0.070

Table 9.9: Conditional probabilities computed by Rules 9.1.6 to 9.1.11, for the proposition 'ultrasound_left_adnexa in { enlarged, mass, cyst }'. (Note that Rule 9.1.7 always fires because its antecedent is 'true'.) Also shown are the relative frequencies with which the conclusion holds given each pattern over the training set of 762 cases.

Pattern of Rule Firing						Computed	Relative
9.1.6	9.1.7	9.1.8	9.1.9	9.1.10	9.1.11	Probability	Frequency
0	1	0	0	0	0	0.025	15/666 = 0.023
0	1	0	0	0	1	0.057	0/ 0
0	1	0	0	1	0	0.624	7/ 10 = 0.700
0	1	0	0	1	1	0.799	0/ 0
0	1	0	1	0	0	0.742	35/ 45 = 0.778
0	1	0	1	0	1	0.873	9/ 12 = 0.750
0	1	0	1	1	0	0.995	1/ 1 = 1.000
0	1	0	1	1	1	0.998	0/ 0
0	1	1	0	0	0	0.183	2/ 16 = 0.125
0	1	1	0	0	1	0.349	3/ 5 = 0.600
0	1	1	0	1	0	0.936	0/ 0
0	1	1	0	1	1	0.972	0/ 0
0	1	1	1	0	0	0.962	2/ 2 = 1.000
0	1	1	1	0	1	0.984	1/ 1 = 1.000
0	1	1	1	1	0	0.999	0/ 0
0	1	1	1	1	1	1.000	0/ 0
1	1	0	0	0	0	0.000	0/ 4 = 0.000
1	1	0	0	0	1	0.000	0/ 0
1	1	0	0	1	0	0.000	0/ 0
1	1	0	0	1	1	0.000	0/ 0
1	1	0	1	0	0	0.000	0/ 0
1	1	0	1	0	1	0.000	0/ 0
1	1	0	1	1	0	0.000	0/ 0
1	1	0	1	1	1	0.000	0/ 0
1	1	1	0	0	0	0.000	0/ 0
1	1	1	0	0	1	0.000	0/ 0
1	1	1	0	1	0	0.000	0/ 0
1	1	1	0	1	1	0.000	0/ 0
1	1	1	1	0	0	0.000	0/ 0
1	1	1	1	0	1	0.000	0/ 0
1	1	1	1	1	0	0.000	0/ 0
1	1	1	1	1	1	0.000	0/ 0
						(total)	35/762 = 0.046

Table 9.10: Variables in the causal rule-based system which fail the χ^2 test at the 1% significance level. Cross-references are given to tables showing the expected and actual frequency distributions.

Variable	χ^2	1% Threshold	Cross Reference
final diagnosis	75.26	34.80	Table 9.11
pain is aggravated by movement	7.42	6.63	
pain is relieved by lying still	7.67	6.63	
progress of pain	11.97	11.30	
type of bleeding	16.34	9.21	
progress of bleeding	19.94	11.30	
contraception	368.60	20.10	
time since appendicectomy	138.71	13.30	
time since laparoscopy	71.86	13.30	
time since laparotomy	32.77	13.30	
previous cervical surgery	1044.92	6.63	Table 9.12
time since cervical surgery	108.30	13.30	
time since tubal ligation	27.26	13.30	
time since right oophorectomy	55.67	13.30	
time since left oophorectomy	34.75	13.30	
time since left salpingectomy	39.66	13.30	
time since Caesarian section	217.14	13.30	
time since hysterectomy	15.39	13.30	
time since termination	14.39	13.30	
time since D+C	30.29	13.30	
speculum blood	18.24	9.21	Table 9.13
pregnancy test	18.51	13.30	
urine microscopy red cells	10.75	9.21	
ultrasound uterine cavity	27.81	20.10	
pregnancy since LMP	92.85	16.80	Table 9.14
raised progesterone	544.16	11.30	
left ovarian cyst	43.50	15.10	
right ovarian cyst	40.01	15.10	

Table 9.11: A comparison of the actual frequency distribution of the variable 'final diagnosis' with that expected according to the causal rule-based system, amongst all 1270 cases.

Value	Predicted Probability	Expected Frequency	Actual Frequency
non specific pain	0.2457	312.0	253
threatened abortion	0.0619	78.6	95
abortion	0.3333	423.3	468
retained products	0.0345	43.8	48
hydatidiform mole	0.0040	5.1	4
ectopic pregnancy	0.0527	66.9	72
pelvic inflammatory disease	0.1409	179.0	149
ovarian cyst	0.0189	24.0	39
cystic accident	0.0375	47.6	54
pelvic haematoma	0.0009	1.1	5
fibroids	0.0033	4.2	8
hyperstimulation	0.0006	0.8	3
urinary tract infection	0.0157	19.9	12
endometriosis	0.0085	10.8	20
ureteric colic	0.0031	3.9	1
appendicitis	0.0022	2.8	2
hyperemesis gravidarum	0.0135	17.2	14
abdominal wall haematoma	0.0013	1.7	1
other	0.0215	27.3	22
(total)	1.0000	1270.0	1270

Table 9.12: A comparison of the actual frequency distribution of the variable 'previous cervical surgery' with that expected according to the causal rule-based system, amongst all 1270 cases.

Value	Predicted Probability	Expected Frequency	Actual Frequency
false	0.5000	635.0	1211
true	0.5000	635.0	59
(total)	1.0000	1270.0	1270

Table 9.13: A comparison of the actual frequency distribution of the variable 'speculum blood' with that expected according to the causal rule-based system, amongst the 1015 cases in which this variable was recorded.

Value	Predicted Probability	Expected Frequency	Actual Frequency
false	0.4730	480.1	443
blood	0.4735	480.6	489
products	0.0535	54.3	83
(total)	1.0000	1015.0	1015

Table 9.14: A comparison of the actnal frequency distribution of the variable 'pregnancy since LMP' with that expected according to the causal rule-based system, amongst all 1270 cases.

Value	Predicted Probability	Expected Frequency	Actual Frequency
false	0.3938	500.1	452
uterine	0.5454	692.7	738
left tubal	0.0124	15.7	35
right tubal	0.0122	15.5	36
left ovarian	0.0162	20.6	1
right ovarian	0.0156	19.8	2
hydatidiform mole	0.0044	5.6	6
(total)	1.0000	1270.0	1270

Only six other variables exceed the 1% threshold by any significant margin: 'final diagnosis', 'type of bleeding', 'progress of bleeding', 'speculum blood', 'pregnancy test', and 'ultrasound uterine cavity'. Of these, the worst two are 'final diagnosis' and 'speculum blood'. The expected and actual distributions for 'final diagnosis' are shown in Table 9.11. The most significant discrepancy occurs with the rarer conditions (pelvic haematoma, fibroids, hyperstimulation, endometriosis, ureteric colic), in which (except for endometriosis) there are only a few positive examples in the database. The most significant discrepancy in the case of 'speculum blood' is underestimation of the frequency with which products of conception are observed: see Table 9.13. However, the marginal distribntions of these variables are not wholly unrealistic, as the tables show. Furthermore, we have also shown that the model generates cases which an expert observer cannot distinguish from real cases [Tod93b]. This experiment was performed part way through the programme of data collection, so only the first 500 of the 1270 cases were available at that time for training purposes. The experiment has not been repeated since because it required a considerable effort on the part of the expert subject: he had to consider carefully 200 cases.

9.2 Inferential Models

9.2.1 Flowchart

Our experience with writing the flowchart was that the ternary decision structure ('T', 'F' and '?') encourages a top-down approach. Near the root of the chart, almost any expression can be used, no matter how abstract the concept that it represents. This is because it is not necessary to be told the truth value of such expressions when using the chart to diagnose cases: the '?' branch is simply followed, and then application of the chart iterates. The freedom to use highly discriminative abstract expressions high in the chart makes it much easier to structure the chart in a clean, logical way. The only constraint is that it should be feasible to write reliable subcharts which can later be attached to the 'unknown' outcomes to determine the truth values of the decision expressions. For example, in practice it will seldom be known *a priori* whether fluid is in fact present in the peritoneal cavity. Nevertheless, we can use this useful discriminant because we are able to attach a subchart to the 'unknown' branch which serves to determine whether fluid is present. We applied these principles recursively as we wrote the subcharts, introducing new additional variables (e.g. 'uterine contents') whenever it was convenient to do so. So that the tree would remain reasonably balanced, as we moved deeper into the chart we selected decision expressions whose truth values were more readily determinable (i.e. only small subcharts need be attached to the 'unknown' outcome.). Thus we start at the root with a decision as to whether the patient became pregnant since the LMP. If so, we then decide whether the pregnancy is (or was) uterine or ectopic. If the pregnancy was uterine we then decide whether it is still viable. (See Figure 9.6.) Notice how the decisions become progressively more concrete, until we are at last able to make a diagnosis.

As a result of this top-down approach, the chart tends to be stable during development. For example, in our case, the chart reflects a hierarchical classification of disorders based on their causal mechanisms, and this is unlikely to require wholesale revision. Alterations tend to be local; only twice did we delete sections, and these both involved only a few nodes. After writing the chart, we tried it on 13 published cases [Gil91], and on 49 cases that had been supplied to us from another centre. As a result we identified six errors in the chart:

Figure 9.6: Nodes at the root of the flowchart. See Table 9.15 for brief descriptions of subcharts $C_1 \ldots C_9$. (Note that subchart C_5 is shown in Figure 9.7, and subchart C_7 is shown in Figure 9.8.)

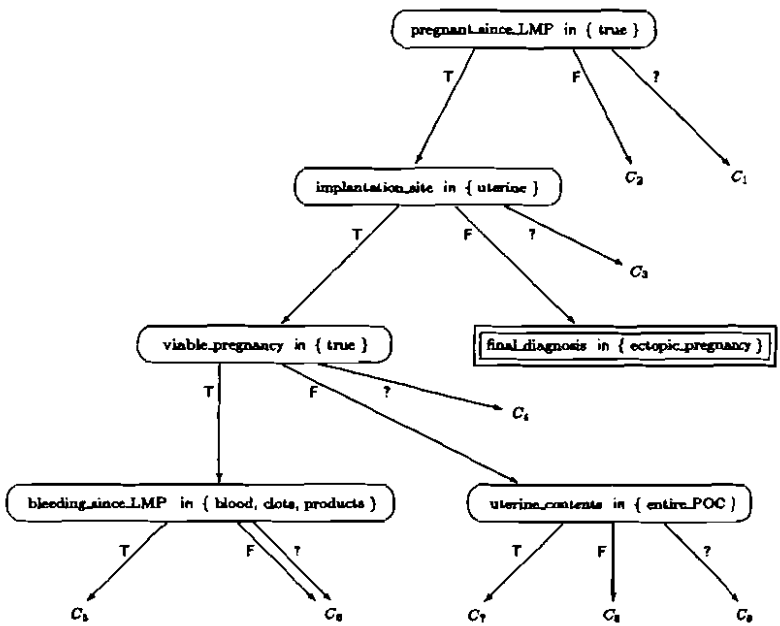


Figure 9.7: Subchart C_5. (See Figure 9.6.)

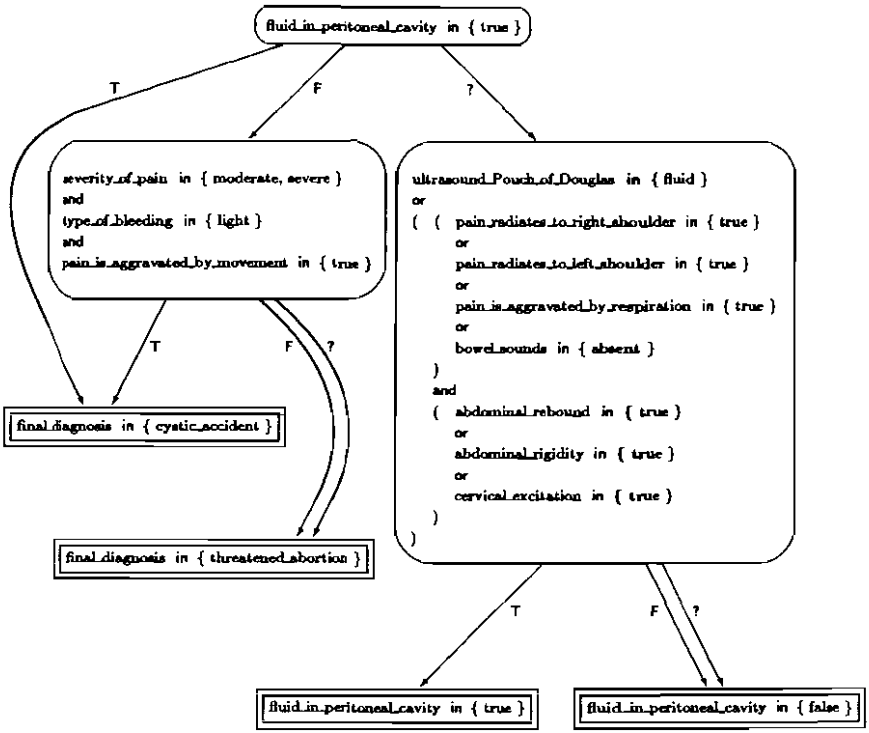


Figure 9.8: Subchart C_7. (See Figure 9.6.)

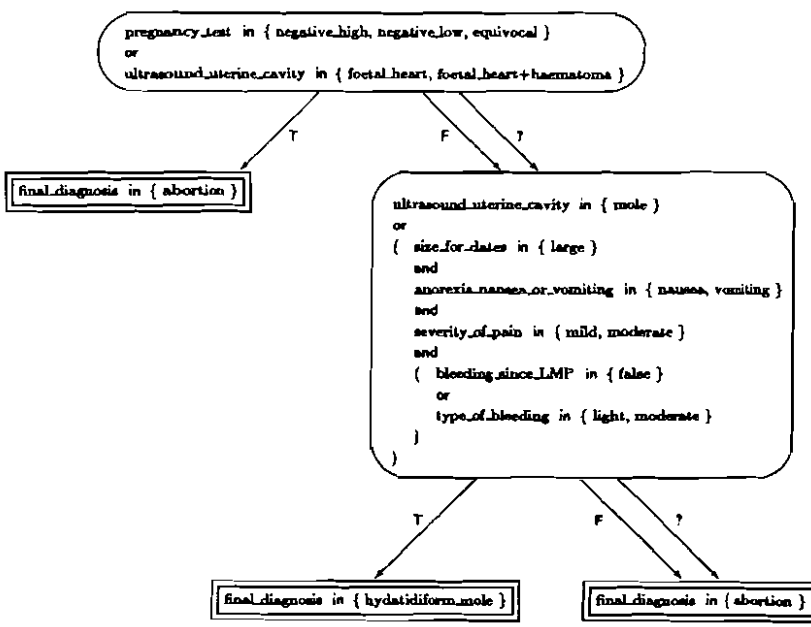


Table 9.15: Summaries of subcharts $C_1 \ldots C_9$ shown in Figure 9.6.

Subchart	Summary
C_1	Determines whether the patient became pregnant since the last menstrual period.
C_2	Determines the final diagnosis in the case that the patient is known not to have become pregnant since the last menstrual period.
C_3	Determines whether the pregnancy is (or was) uterine or ectopic in the case that the patient is known to have become pregnant since the last menstrual period.
C_4	Determines whether the pregnancy is currently viable in the case that the patient is known to have become pregnant *in utero* since the last menstrual period.
C_5	Determines whether the final diagnosis is 'cystic accident' or 'threatened abortion' in the case that abnormal bleeding is reported by a patient who is known to have a viable uterine pregnancy. (See Figure 9.7.)
C_6	Determines the final diagnosis in the case that the patient is known to have a viable uterine pregnancy, and reports no abnormal bleeding.
C_7	Determines whether the final diagnosis is 'hydatidiform mole' or 'abortion' in the case that the patient is known to have an unviable uterine pregnancy, and all products of conception are *in utero*. (See Figure 9.8.)
C_8	Determines the final diagnosis in the case that the patient is known to have become pregnant since the LMP, but not all prodncts of conception are still *in utero*.
C_9	Determines whether all products of conception are still *in utero* in the case that the patient is known to have become pregnant since the LMP, but the pregnancy is no longer viable.

three were typing errors and three were oversights. They all required only trivial changes. Finally we tried the chart on the first 202 cases in the database. The flowchart made the correct diagnosis in 119 (58.9%) of these. We traced the decisions for all cases that the flowchart misdiagnosed, and if this highlighted a specific weakness we modified the chart. As a result, the flowchart then correctly diagnosed 142 (70.3%) of the same 202 cases.

The use of abstract discriminants encourages their reuse in other parts of the chart. Charts therefore have compact representations as acyclic graphs. Our flowchart has 101 non-terminal nodes when represented as an acyclic graph, and it references 101 variables in all. Leaf expressions are invariably simple propositions in the form of assertions that a particular variable has a single value.

9.2.2 Inferential Rule-Based System

One of the difficulties we encountered in constructing an inferential rule-based system was in identifying pathophysiological states whose presence can be determined reliably from observations without necessarily knowing the final diagnosis. We finally selected nine additional variables that seemed suitable, and supplemented them with four new variables that we had used as subgoals when constructing the flowchart program. See Table 9.16.

Table 9.16: The 13 additional variables used in the inferential rule-based system. (The variables derived from the flowchart are marked with a '$\star$'.)

recent TOP or ERPC ($\star$)
recent previous abortion ($\star$)
uterine contents ($\star$)
viable pregnancy ($\star$)
pregnancy since LMP
threatened abortion
left ovarian cyst
right ovarian cyst
microscopic haematuria
fibroids
acute red degeneration
peritoneal cavity
hyperstimulation

Since symptom variables are often unrecorded, it is extremely unlikely in any given case that the atomic propositions whose truth values are observed will form an initial segment of the sequence Q, whichever Q we choose. We therefore elected to treat unrecorded symptom variables as having the explicit value 'unknown'. This means that the truth value of any atomic proposition involving a symptom value is always known. Therefore, provided that no symptom variable is preceded by an additional variable (or the final diagnosis!) in the sequence Q, the set of propositions whose truth value is given will necessarily form

(the same) initial segment of Q. Since the families of rules concerning those propositions are therefore redundant, we naturally decided to omit all propositions involving symptom variables from our sequence. We refer to this sequence as Q_I because it has an inferential ordering. Thus Q_I is incomplete containing only 35 propositions (c.f. Q_C which has 478 propositions) relating to the 13 additional variables and the final diagnosis. The rules associated with these propositions of course include symptom variables in their antecedents, although we included only 106 of the 169 symptom variables which we regarded as the most useful diagnostically.

Sequence Q_I begins with propositions concerning pathophysiological states that are most immediately diagnosable, such as pregnancy and peritonitis. Propositions concerning the anatomical site of the pregnancy, the viability of the pregnancy, and the nature of the uterine contents, follow in that order. Ectopic pregnancy has various specific risk factors which are readily enumerated, and of course requires that the patient is pregnant. Therefore, since an ovarian cyst is sometimes difficult to distinguish from an ectopic pregnancy, a decision as to the presence of an ovarian cyst is left until after that of ectopic pregnancy. Determination of the precise cause of the pain (the 'final diagnosis') is the last task, and this involves the last 18 propositions. The average number of rules associated with each proposition is between six and seven (slightly more than in the causal rule-based system): there are a total of 221 rules in the knowledge base, of which 120 are categorical (84 logical preclusions and 36 logical implications).

For example, shown below is the family of rules we formulated to determine whether the final diagnosis is 'urinary tract infection'. There are seven rules altogether. The certainty factors shown are those derived by iterative optimization (except for Rule 9.2.1 which is categorical) using the entire database of 1270 cases as a training set.

Rule 9.2.1

not final_diagnosis in {non_specific_pain,
ovarian_cyst,
urinary_tract_infection,
endometriosis,
acute_appendicitis,
abdominal_wall_haematoma,
other }
$\Rightarrow^{0}$ final_diagnosis in { urinary_tract_infection }

This first rule is categorical. It dictates that if the final diagnosis has already been established as something else by anterior rules, then the final diagnosis cannot be urinary tract infection. This rule is required in order to ensure external consistency (Equation 6.14). It is slightly simpler to express this rule in the form 'unless the final diagnosis is one of the remaining possibilities'.

Rule 9.2.2

true $\Rightarrow^{0.016}$ final_diagnosis in { urinary_tract_infection }

This rule reflects the relatively low prevalence of urinary tract infections amongst our group of patients.

Rule 9.2.3

```
viable_pregnancy in { true }
or
past_history_of_UTI in { true }
    ⇒0.597 final_diagnosis in { urinary_tract_infection }
```

This rule identifies two risk factors for urinary tract infection. A previous urinary infection makes a subsequent one slightly more likely, and urinary infections are commoner in pregnancy owing to urinary stasis.

Rule 9.2.4

```
frequency in { true }
and
( pregnancy_since_LMP in { false }
  or
  dysuria in { true }
)
    ⇒0.707 final_diagnosis in { urinary_tract_infection }
```

The fourth rule refers to the specific symptoms of a urinary tract infection: urinary frequency and pain on passing urine ('dysuria'). Pregnancy is also a cause of urinary frequency, and so we guard frequency by the proviso that the patient is not pregnant.

Rule 9.2.5

```
site_of_tenderness in { left_loin, right_loin }
    ⇒0.745 final_diagnosis in { urinary_tract_infection }
```

The fifth rule refers to the signs of kidney infection: loin tenderness.

Rule 9.2.6

```
recent_fever_or_chill in { true }
or
temperature in { 37.5_to_38.0, 38_or_more }
or
white_cell_count in { 11.0_or_more }
    ⇒0.546 final_diagnosis in { urinary_tract_infection }
```

The sixth rule refers to general symptoms and signs of infection: recent fevers or chills, an elevated temperature, and a raised white cell count. We do not discriminate in this rule between mildly elevated temperatures (37.5° C to 38.0° C) and higher temperatures (over 38.0° C), although we would do so if sufficient training data were available.

Rule 9.2.7

microscopic_haematuria in { true }
or
(urine_microscopy_pus_cells in { moderate }
and
urine_microscopy_squames in { none, minimal }
)
$\Rightarrow^{0.877}$ final_diagnosis in { urinary_tract_infection }

The seventh and last rule refers to evidence of red blood cells or pus cells in the urine. The latter are not significant if there are more than a few squamous cells present because squames indicate external contamination at the time of collecting the specimen.

The certainty factors of the 101 non-categorical rules were derived from the given training sample by iterative maximum likelihood estimation, just as for the causal rule-based system. Figure 9.9 plots mean surprise as a function of training iteration for the family of rules shown above: since the final diagnosis is always recorded, the truth value of the proposition 'final_diagnosis in { urinary_tract_infection }' was known in all 1270 cases. Clearly there is no significant reduction in surprise after about 500 iterations, and 1000 iterations would therefore appear to be adequate, as it was for the causal rule-based system (Figure 9.5).

Table 9.17 shows the frequency distribution of all 101 non-categorical certainty factors. The distribution is roughly uniform, as it is for the causal rule-based system (Table 9.7). Table 9.18 enumerates the conditional probabilities defined by the family of rules shown above. As in the case of the causal rule-based system, although there are roughly three times as many data as there are degrees of freedom (rules with non-categorical certainty factors), a reasonable fit is obtained.

Figure 9.9: Graph of average surprise per training case as a function of training iteration for the family of rules in the inferential rule-based system.

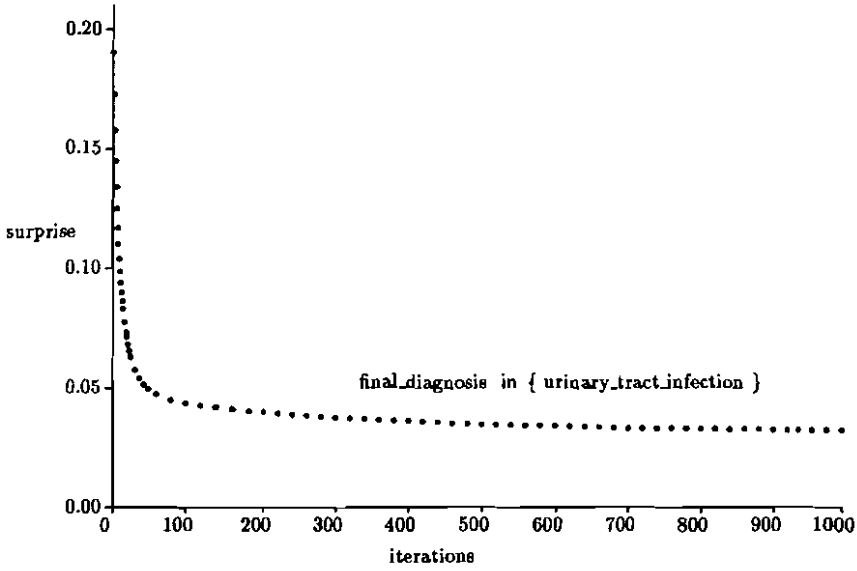


Table 9.17: Frequency distribution of the 101 non-categorical certainty factors in the inferential rule-based system derived from the entire database of 1270 cases.

Interval	Frequency
(0.0, 0.1]	16
(0.1, 0.2]	7
(0.2, 0.3]	10
(0.3, 0.4]	12
(0.4, 0.5)	1
[0.5, 0.6)	8
[0.6, 0.7)	10
[0.7, 0.8)	14
[0.8, 0.9)	9
[0.9, 1.0)	14
(total)	101

Table 9.18: Conditional probabilities of 'final_diagnosis in { urinary_tract_infection }' computed by the corresponding family of rules: Rules 9.2.1 to 9.2.7. Note that Rule 9.2.2 always fires because its antecedent is 'true'. For brevity, all patterns in which Rule 9.2.1 fires are pooled, because when it fires, the computed probability is always zero, and no actual examples are found. Also shown are the relative frequencies with which the conclusion holds given each pattern over the training set of 1270 cases.

Pattern of Rule Firing							Computed	Relative Frequency
9.2.1	9.2.2	9.2.3	9.2.4	9.2.5	9.2.6	9.2.7	Probability	
0	1	0	0	0	0	0	0.016	1/ 152 = 0.007
0	1	0	0	0	0	1	0.106	1/ 7 = 0.143
0	1	0	0	0	1	0	0.020	0/ 36 = 0.000
0	1	0	0	0	1	1	0.125	0/ 4 = 0.000
0	1	0	0	1	0	0	0.046	1/ 2 = 0.500
0	1	0	0	1	0	1	0.257	0/ 0
0	1	0	0	1	1	0	0.055	0/ 1 = 0.000
0	1	0	0	1	1	1	0.294	0/ 0
0	1	0	1	0	0	0	0.039	2/ 39 = 0.051
0	1	0	1	0	0	1	0.223	1/ 5 = 0.200
0	1	0	1	0	1	0	0.046	1/ 13 = 0.077
0	1	0	1	0	1	1	0.256	0/ 2 = 0.000
0	1	0	1	1	0	0	0.105	0/ 1 = 0.000
0	1	0	1	1	0	1	0.455	0/ 0
0	1	0	1	1	1	0	0.124	0/ 0
0	1	0	1	1	1	1	0.502	0/ 0
0	1	1	0	0	0	0	0.024	1/ 54 = 0.019
0	1	1	0	0	0	1	0.149	0/ 3 = 0.000
0	1	1	0	0	1	0	0.029	0/ 14 = 0.000
0	1	1	0	0	1	1	0.174	1/ 1 = 1.000
0	1	1	0	1	0	0	0.067	0/ 0
0	1	1	0	1	0	1	0.339	1/ 2 = 0.500
0	1	1	0	1	1	0	0.080	0/ 0
0	1	1	0	1	1	1	0.381	0/ 0
0	1	1	1	0	0	0	0.056	0/ 9 = 0.000
0	1	1	1	0	0	1	0.298	0/ 1 = 0.000
0	1	1	1	0	1	0	0.067	0/ 1 = 0.000
0	1	1	1	0	1	1	0.338	2/ 2 = 1.000
0	1	1	1	1	0	0	0.148	0/ 0
0	1	1	1	1	0	1	0.553	0/ 0
0	1	1	1	1	1	0	0.173	0/ 0
0	1	1	1	1	1	1	0.598	0/ 0
1	1	-	-	-	-	-	0.000	0/ 921 = 0.000
							(total)	12/1270 = 0.009

Chapter 10

Evaluation

In this chapter we describe how we have trained and tested the diagnostic programs, and we present our results. First we explain how we chose the domain-specific parameters (α, k, etc.) for the various methods. We then present the diagnostic accuracies of all methods obtained using cross-validation on batches of 101 test cases. Knowledge-based methods were found to be no more accurate than the best statistical methods. The latter include both the nearest neighbours method using the Bayes metric and independence Bayes itself. The experiments were repeated using only the cases in the database with definite final diagnoses, and similar results were obtained. Finally, nearest neighbours was compared with independence Bayes on all cases using a leave-out-one cross-validation strategy. No significant difference in accuracy could be demonstrated. It is argued that nearest neighbours using the Bayes metric is one of the most accurate methods and the most suitable technique for providing machine assistance for medical diagnosis.

10.1 Training and Testing

In total we have available 1270 cases in the database. However, the first 202 cases (set 'A') cannot be used as test cases for the knowledge-based methods because these cases were used to assist construction of the knowledge-bases. Therefore, since we wished to compare all programs on an identical test set initially, only the remaining 1068 cases (set 'B') were available for testing. The first 202 cases (set A) were available for training. Furthermore, in order to increase the number of training cases available, we adopted a cross-validation strategy; for training purposes we including cases from set B as well. However, some methods (neural networks and the rule-based system with a causal ordering) are so expensive to train that it was not feasible to leave out only the test case because this would have meant retraining the classifier 1068 times. Instead, as a compromise, we chose to leave out 101 cases. Also, since some patients appear more than once in the database, and since repeat presentations of the same patient tend to resemble the previous one, we also omitted from the training set any cases which represented the same patient as any of the 101 cases in the test batch. We therefore partitioned set B into 11 subsets ($B = B_1 \cup B_2 \cup \ldots \cup B_{11}$) and before testing on each subset B_i, we trained the classifier on set A and every other subset set B_j ($j \neq i$) with the exception of any other presentations of patients in B_j. Table 10.1

shows the actual size of each training and test set. Notice that although on average each patient appears 1.3 times in the database, repeat presentations tend to be within the same test set and so do not require removal from the training set: on each row of Table 10.1, the test set and training set together contain nearly 1270 cases. This is because multiple presentations of the same patient during the study period tended to be entered into the database sequentially since all such presentations were usually recorded in the same case-notes.

Table 10.1: Sizes of each training and test set.

Test set	Size of test set	Size of training set
B_1	101	1164
B_2	101	1164
B_3	101	1168
B_4	101	1169
B_5	101	1169
B_6	101	1168
B_7	101	1167
B_8	101	1168
B_9	101	1169
B_{10}	101	1165
B_{11}	58	1210
Total (B)	1068	-

For each test case, we took the disease with highest posterior probability as the computer's diagnosis. This was counted as an error if it disagreed with the recorded diagnosis for the case. Table 10.4 (Page 98) shows the overall error rates for each program. Some diagnostic programs require selection of appropriate training algorithms and/or choice of domain-specific parameters. We describe these for each method in turn.

10.1.1 Independence Bayes

One possible tunable parameter for independence Bayes is the number of symptom variables that are actually used. Crichton *et al* [Cri87, Cri89] report a small increase in accuracy using 5 or 6 selected variables rather than all 41 on a database of acute abdominal pain, though the improvement does not appear to be statistically significant. In the diagnosis of acute coronary heart disease by means of independence Bayes, Aase *et al* [Aas93] found an improvement if only 31 instead of all 38 variables were used. However, the quadratic score rather than diagnostic accuracy was used as the performance measure: elimination of dependent variables would be expected to improve calibration, and hence could conceivably improve the quadratic score at the expense of the diagnostic accuracy. Furthermore, a reclassification estimate of the quadratic score appears to have been used: it does not follow necessarily that the quadratic score on unseen cases would be higher with 31 variables than with all 38. Ohmann *et al* [Ohm86] warned of the danger of bias in selecting variables,

and in their study of patients with gastrointestinal bleeding, they found that the estimated true diagnostic accuracy generally improved as variables were included. Similarly, in a simulation study of cases of vaginal discharge, Chard and Rubenstein [Cha89] found that optimal diagnostic accuracy was achieved using all variables. Furthermore, although equal overall accuracy could apparently be achieved using fewer variables, the ability to correctly detect less prevalent conditions was impaired. For these reasons, and because on average only 62% of variables were recorded in each case and so it is not clear *a priori* which variables to select, we opted to include *all* symptom variables in our implementation of independence Bayes.

10.1.2 Nearest Neighbours

The relevant parameter here is k, the number of neighbours from which to draw statistical inference. If k is too large, then the neighbours are not representative of the test case. If k is too small then random noise degrades the accuracy of classification. Clearly optimal choice of k depends on the domain of application and the size of the training set. Figure 10.1 shows graphs of error rate obtained with the two metrics as a function of k. The error rate for the Hamming metric is significantly higher than for the Bayes metric for a wide range of k values. The error rates shown in Table 10.4, 0.485 (Hamming metric) and 0.362 (Bayes metric), are for optimal k chosen retrospectively ($k = 21$ and $k = 19$, respectively). The horizontal nature of the graphs suggests that similar error rates would be obtained using the same k values if a further random sample of cases were used.

10.1.3 Iterative partitioning

The relevant parameter for iterative partitioning is the stopping threshold α. If α is small then partitioning stops soon and so the filtered subset of training cases is large and unrepresentative of the test case. If α is large then the filtered subset is small and random noise degrades the accuracy of classification. Figure 10.2 shows a graph of error rate for iterative partitioning as a function of α. Notice that $\alpha = 28.9$ and $\alpha = 34.8$ correspond to a 95% and 99% significance threshold, respectively. This is because the likelihood ratio statistic (Equation 4.22), when reduced by a factor of 2, approximates a Chi-Square distribution, and the 95% and 99% significance thresholds for the Chi-Square distribution with 18 degrees of freedom (#Disease − 1) are 14.45 and 17.4, respectively.

The error rate shown in Table 10.4 (0.417) is for optimal retrospective choice of α (26.0). Again the horizontal nature of the graph suggests that a similar error rate would be obtained with the same α value if a further random sample of test cases were used. With $\alpha = 26.0$, surprisingly few facts about the test case are actually used in diagnosis: on average the database is partitioned only 1.88 times (i.e. 1.88 facts are used). The actual number of iterations varies from 1 to 4 (Table 10.2). This means that the filtered subset of the training set is large: on average it contains 129.5 cases (standard deviation 78.6 cases). Figure 10.3 shows a graph of the cumulative frequency of sizes of filtered subsets.

10.1.4 Neural network

The training method we adopted for neural networks was *back-propagation* [Rum86a] minimising the mean squared error. Although new optimization algorithms are frequently

Figure 10.1: Graphs of error rate for the nearest neighbours method as a function of k for the Hamming metric and Bayes Metric.

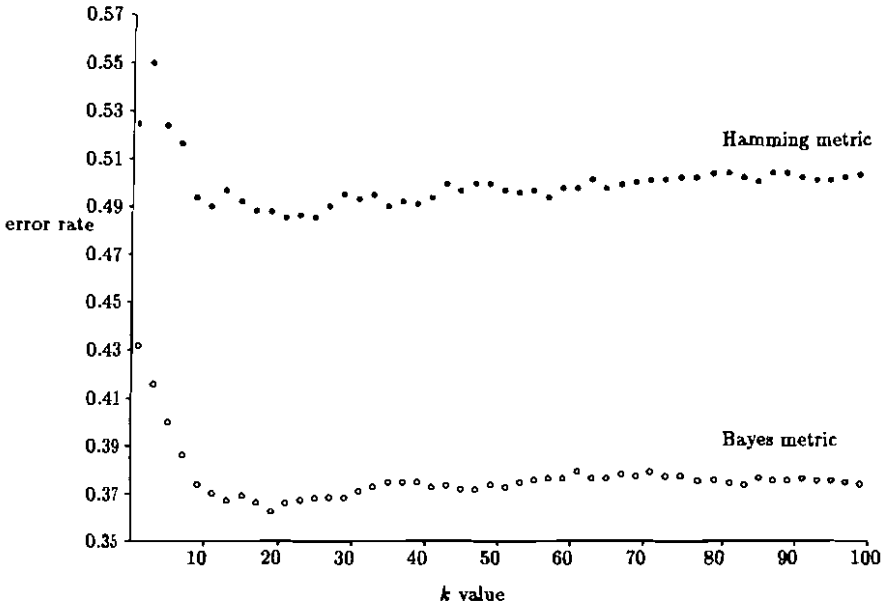


Table 10.2: Frequency distribution of the number of iterations performed by the iterative partitioning program with respect to the 1068 test cases.

Iterations	Frequency
1	184
2	829
3	54
4	1
Total	1068

Figure 10.2: Graph of error rate for iterative partitioning as a function of α. Also shown are the corresponding 95% ($\alpha = 28.9$) and 99% ($\alpha = 34.8$) significance thresholds.

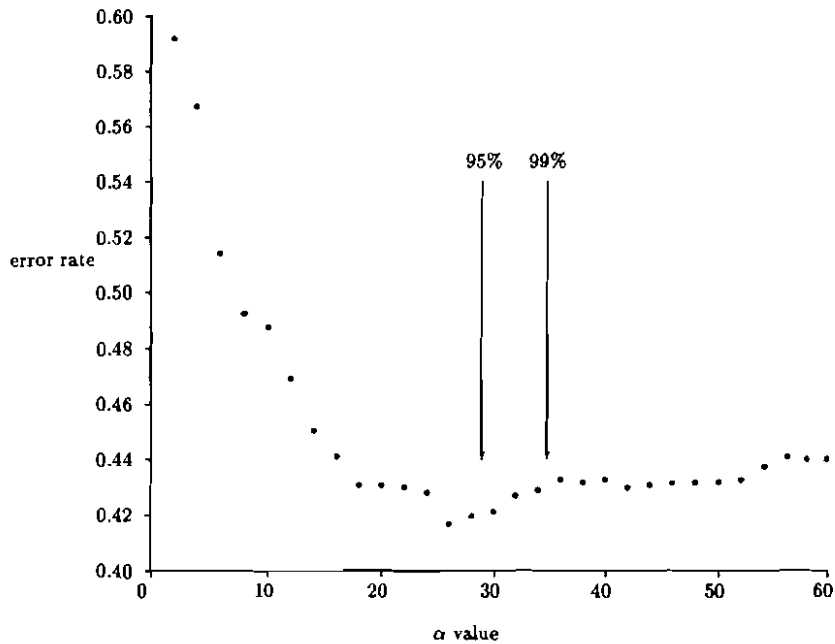
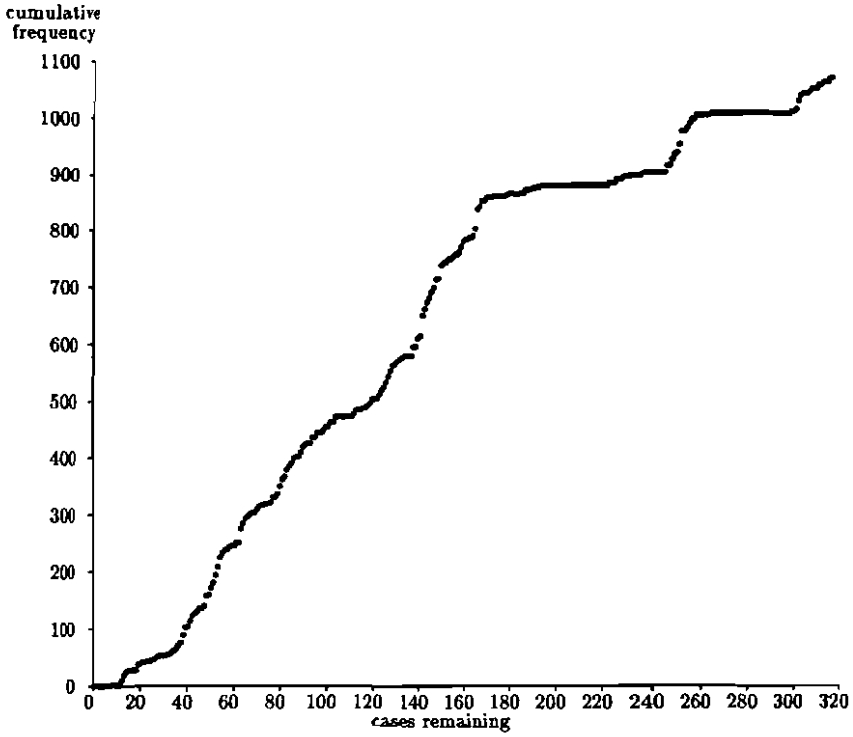


Figure 10.3: Graph for iterative partitioning plotting the cumulative frequency distribution of sizes of filtered subsets of the training set with respect to the 1068 test cases.



proposed, back-propagation remains a simple and effective method; in a recent comparative study of three training algorithms, back-propagation was found to be the best [Ebe91]. In order to break symmetry we set all weights initially to small random values (drawn uniformly from the interval $[-0.2, 0.2]$). Owing to the large number of weights, and the relatively large number of training cases, it was too expensive to calculate the total error over the entire training set at each iteration. Therefore we applied the variation of back-propagation in which weights are modified after each presentation of a single training case. The training cases were presented in fixed order, and 100 passes were made through the training set. In order to achieve convergence, the gain ('learning rate') was reduced linearly to zero. No momentum term was employed. Two parameters remained to be determined for optimum performance: the initial gain, and the number of hidden units. We tried various combinations of these; Table 10.3 shows the corresponding error rates. The lowest error rate obtained was 0.378 (404 errors) with one hidden unit and an initial gain of 0.03. This is the error rate shown in Table 10.4. Figure 10.4 plots error rate as a function of the total number of passes through the training set (initial gain 0.03, 1 hidden unit). Clearly no further improvement in error rate could be expected had we employed more than 100 passes. It is perhaps not surprising that increasing the number of hidden units has little effect on error rate: the network already has a large number of degrees of freedom with so many direct inputs (554) to each of the 19 output units.

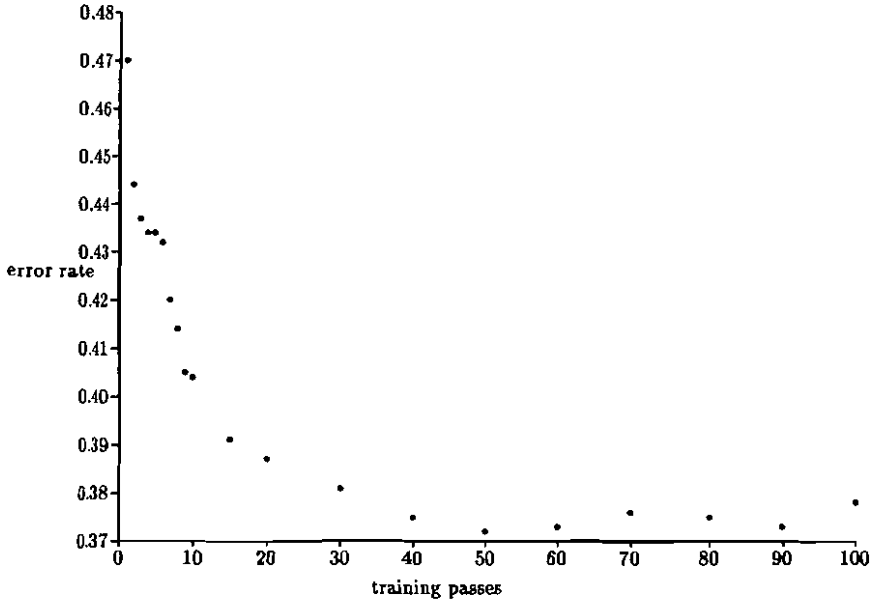
Table 10.3: Error rates of the neural network with respect to the 1068 test cases when using different numbers of hidden units n and various initial gains.

	Numbers of Hidden Units									
	$n = 0$		$n = 1$		$n = 2$		$n = 5$		$n = 10$	
Initial Gain	Total Errors	Error Rate	Total Errors	Error Rate	Total Errors	Error Rate	Total Errors	Error Rate	Total Errors	Error Rate
0.003	443	0.415	431	0.404	450	0.421	442	0.414	439	0.411
0.01	413	0.387	410	0.384	411	0.385	407	0.381	410	0.384
0.03	410	0.384	404	0.378	415	0.389	408	0.382	405	0.379
0.1	430	0.403	423	0.396	419	0.392	416	0.390	418	0.391
0.3	427	0.400	422	0.395	433	0.405	430	0.403	442	0.414
1.0	442	0.414	450	0.421	447	0.419	443	0.415	442	0.414

10.1.5 Causal Rule-Based System

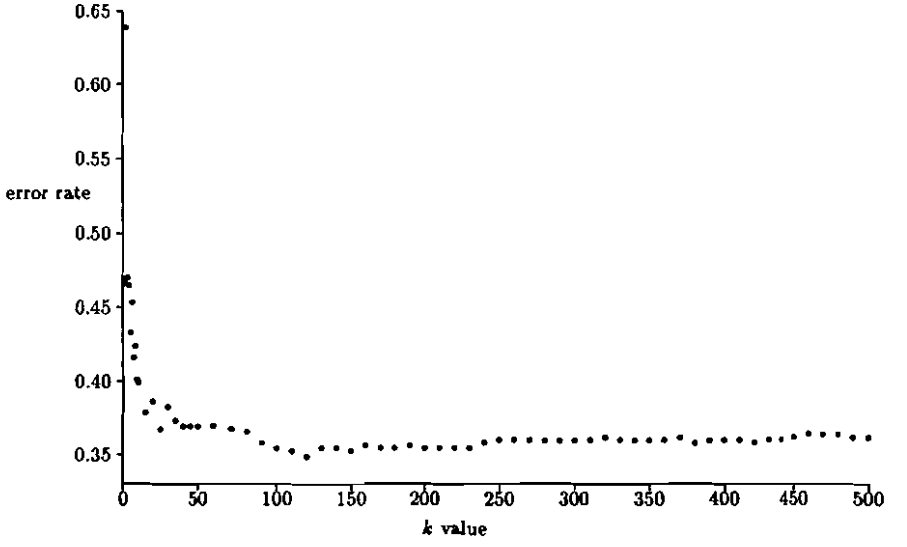
Regarding the causal rule-based system, it was not feasible to apply the exact inference algorithm described by Lauritzen and Spiegelhalter [Lau88] because, viewed as a Bayesian network over binary propositions, 27 of the nodes (atomic propositions) have more than 20 parents, the worst having 42 parents. Therefore, we employed the simple Monte Carlo method described in Chapter 6; we generated a large sample (10^6) of random simulated cases, and using these as a training sample for a statistical method. The statistical classifier we employed was nearest neighbours with the Bayes metric since this was found to be the most accurate of all the statistical methods we had tried (Table 10.4). Since we were using a 'leave-out-101' strategy, the rule-based system had to be trained on the appropriate set

Figure 10.4: Graph plotting error rate for the neural network on the 1068 test cases as a function of the total number of passes through the corresponding training sets ('leave-out-101').



of real cases, and then used to generate the simulated cases afresh for each batch of test cases. The simulated cases were then used first to derive the parameters for the Bayes metric, and then as a reference database from which to extract the nearest neighbours to each real test case. The accuracy of the method clearly depends on k, the number of neighbours extracted. Figure 10.5 shows a graph of error rate using this method on all 1068 cases as a function of the number of neighbours. Retrospectively, a k value of 100 is about optimal, producing an error rate of 0.352, and this is the value entered in Table 10.4. For comparison, as a baseline, we repeated this experiment using the independence Bayes classifier itself in place of the nearest neighbours classifier. We obtained exactly the same error rate (0.364) as we did when training on real cases. The two methods did not make identical decisions, however; precisely one of them was correct in 178 of the 1068 cases. The discriminant matrices (Tables B.1 and B.12) are shown in the appendix (Appendix B) for comparison. Although using nearest neighbours rather than independence Bayes as the classifier when training on simulated cases leads to a reduction in error rate (13 fewer errors), this difference is not statistically significant ($p = 0.0854$). The statistical test we use for this and all other comparisons of error rates is the variation on the McNemar test suggested by Mosteller [Mos52]. This treats all cases correctly diagnosed by one method but not the other as a sequence of binomial trials with probability of success $\frac{1}{2}$.

Figure 10.5: Graph of error rate for nearest neighbours with the Bayes metric on all 1068 cases when trained on databases of 10^6 simulated cases generated from the causal rule-based system ('leave-out-101').



10.1.6 Discussion

Table 10.4 shows the error rates for all programs. Full discriminant matrices corresponding to each entry are included in the appendix. The most accurate program is nearest neighbours with the Bayes metric trained on simulated cases generated from the causal rule-based system (376 errors: see Table B.1 for the discriminant matrix). However, this is not significantly better ($p = 0.4671$) than the best purely statistical alternative: nearest neighbours with the Bayes metric trained on real cases (387 errors: see Table B.2 for the discriminant matrix). The most striking difference between the two discriminant matrices is that the sensitivity to non-specific pain (Disease 'A') falls from 65.1% to 54.1% while the specificity rises from 82.2% to 88.8% when simulated cases rather than real cases are used for training. This is despite the fact that the causal rule-based system tends to generate a higher proportion of cases of non-specific pain (0.2457) than are actually observed, 0.1992 ($\approx 253/1270$): see Table 9.11. A similar effect is observed when independence Bayes is used instead of nearest neighbours as the classifier (Tables B.3 and B.12). This suggests that simulated cases of non-specific pain lack the variation of clinical presentation that is actually observed. This is not surprising given the obvious difficulty in modelling a condition that has no clearly understood causal mechanism.

The causal rule-based system is, however, significantly more accurate ($p = 0.0425$) than the inferential rule-based system which makes 408 errors, although this conclusion must be tempered by the fact that the choice of $k = 100$ for the causal system was retrospective. A comparison of the discriminant matrices (Tables B.1 and B.6) shows that the inferential system has a lower threshold for non-specific pain, but is less accurate at diagnosing most other conditions. In particular the inferential rule-based system's diagnoses of the rarer conditions (Diseases 'J' to 'S', inclusive) are overall much less reliable (34.1% compared to 53.6%), yet slightly less sensitive too (18.7% compared to 20.0%). The poorer performance regarding rarer conditions perhaps reflects the small size of the inferential knowledge base compared to the causal one. Certainly when writing the rule-base it was difficult to formulate a small number of rules which would reliably detect the rarer conditions. The flowchart (discriminant matrix Table B.8) is even less accurate (431 errors) than the inferential rule-based system, but again the difference does not reach statistical significance ($p = 0.0788$). The flowchart appears to have an even lower threshold for non-specific pain, and performance regarding the less prevalent conditions (Diseases 'J' to 'S', inclusive) is even worse: reliability 26.9% and sensitivity 9.3%, overall. This can be explained by the greater flexibility of the rule-based system with its numerical certainty factors and its capacity to learn from training examples. Independence Bayes (389 errors) is significantly more accurate than the flowchart ($p = 0.0170$). This reverses an earlier conclusion based on the much smaller sample of 202 training cases [Sta92]. Enlargement of the Bayesian network seems to reduce accuracy. The large Bayesian network makes 20 more errors than independence Bayes, although the difference is not statistically significant ($p = 0.0721$). The general form of the discriminant matrices (Tables B.3, B.4, and B.7) resembles that of the causal rule-based system (Table B.1).

The nearest neighbours program using independence Bayes as a metric makes two fewer errors than independence Bayes itself. A comparison of the two discriminant matrices (Tables B.2 and B.3) shows that nearest neighbours tends to do better with the most common conditions (abortion and non-specific-pain) at the expense of the rarer ones. This is perhaps to be expected with $k = 19$ since there are fewer examples of each rare condition than the

number of neighbours retrieved. The neural network (Table B.5) has a greater sensitivity to common conditions, but is generally poorer than independence Bayes at diagnosing rarer conditions. In particular, the neural network has a greater tendency to misdiagnose cases of pelvic inflammatory disease as having either abortion or non-specific pain. The higher error rate of the neural network (404 errors) compared to independence Bayes (389 errors) is similar to the results of Hart and Wyatt in their study of chest pain [Har89], although Baxt's experience with neural networks for the same problem was very much more rewarding [Bax91]. Phillips *et al* [Phi91] report that a neural network was more accurate than independence Bayes for acute abdominal pain. The number of test cases (30) was small, and the difference does not appear to be statistically significant.

Iterative partitioning (Table B.9) shows an even poorer sensitivity to the less prevalent conditions, no doubt because the average size of the filtered database is even larger (129.5 cases), making 56 more errors than independence Bayes: this difference is highly significant ($p = 0.0017$). The present application does not appear to be particularly well-suited to iterative partitioning, because seldom can a reliable diagnosis be made purely on the basis of just two facts (Table 10.2). Cases of fibroids seem to be an exception. All five cases of fibroids that were correctly identified by iterative partitioning were diagnosed on the strength of the same two facts: no abnormal bleeding since the LMP, and fibroids seen on ultrasound examination. The exemplar model (Table B.10) has particular difficulty with beterogeneous conditions (non-specific pain, ectopic pregnancy, pelvic inflammatory disease and ovarian cyst). This reflects the handicap imposed by having only a single template for each disease. It would be interesting to see how much improvement could be obtained with a more refined model. The worst program is nearest neighbours with the Hamming metric (518 errors). This is strongly biased towards the commoner conditions, diagnosing all but 85 of the 1068 cases as having either non-specific pain or miscarriage.

The error rates we obtained compare favourably with those in other studies using independence Bayes (0.414 [Sut89b],0.457 [Ser86]) when attention is confined to the same range of disorders that we have studied here. Furthermore, the best programs we implemented made fewer errors than the initial clinical diagnosis (error rate 0.399 over the same 1068 cases), althongh the latter should properly be regarded as a lower bound for several reasons:

1. In 43 cases the clinician had not recorded an initial diagnosis, and we counted these as errors.

2. The results of the ultrasound scan may not always have been available when the initial diagnosis was made.

3. Where a list of possible diagnoses was given the first was taken as the initial diagnosis. However, it is quite possible that the clinician enumerated his possible diagnoses in order of decreasing gravity rather than decreasing probability.

10.1.7 CART

As a baseline against which to compare our statistical programs, we also tried the tree-based modelling module provided by the S-Plus statistical package, described more fully in [Cla92]. This implements a version of the well-known CART algorithm first described by Breiman *et al* [Bre84]. A classification tree is built by recursively partitioning the training sample. The features available for partitioning in the S-Plus implementation are atomic

Table 10.4: Error rates for all programs on the test set of 1068 cases using a 'leave-out-101' training strategy, sorted into rank order. The right-hand column lists the corresponding tables in the appendix showing the full discriminant matrices. There is no significant difference between the first five programs at the 5% significance level using the McNemar test.

Program	Errors	Error rate	Matrix
Causal rule-based system (ψ_R, $Q = Q_C$, $T = T_C$)	376	0.352	Table B.1
Nearest Bayes neighbours (ψ_K, $\delta = \delta_b$, $k = 19$)	387	0.362	Table B.2
Independence Bayes (ψ_B, $P = P_I$)	389	0.364	Table B.3
Small Bayesian network (ψ_B, $P = P_S$)	402	0.376	Table B.4
Neural network (ψ_N, $n = 1$)	404	0.378	Table B.5
Inferential rule-based system (ψ_R, $Q = Q_I$, $T = T_I$)	408	0.382	Table B.6
Large Bayesian network (ψ_B, $P = P_L$)	409	0.383	Table B.7
Flowchart (ψ_F)	431	0.404	Table B.8
Iterative partitioning (ψ_I, $\alpha = 26.0$)	445	0.417	Table B.9
Exemplar model (ψ_E)	457	0.428	Table B.10
Nearest Hamming neighbours (ψ_K, $\delta = \delta_J$, $k = 21$)	518	0.485	Table B.11

propositions. The next proposition on which to partition is chosen so as to maximise the decrease in the node deviance (defined as minus twice the log-likelihood of the observations remaining at the node). The stopping criterion used was the default for the S-Plus implementation of CART, namely that a node will not be partitioned if

- the node deviance is less than 1% of the root node deviance, and
- the node has fewer than 10 cases remaining to be partitioned.

This criterion is considered to be quite liberal in [Cla92], resulting in overly large trees which can then be pruned to improve classification accuracy. This strategy is generally considered better than trying to stop tree growth at some optimal point. The method of cost-complexity pruning advocated in [Bre84], and adopted in the S-Plus implementation allocates a cost $D_\alpha(T)$ to a tree T defined by

$$D_\alpha(T) \triangleq D(T) + \alpha \text{size}(T)$$

$D(T)$ is the deviance of the tree, defined as the sum over all leaves of the deviance of that leaf, and the size of a tree is the number of leaves. The value of α determines how heavily larger trees are penalised and thus how much a tree should be pruned. The choice of α can be made by evaluating pruned trees on new data, or if this is not available, by cross-validation. This second approach is supported by the S-Plus implementation and involves splitting the training database into n batches; for each batch in turn a tree is then grown using the remaining $n - 1$ batches, and the sequence of optimal subtrees found as α varies. The withheld batch is then used to calculate the deviance of each of the subtrees, and the

deviances averaged over all batches. This gives an unbiased estimate of the best size of tree to use.

The S-Plus implementation allows missing values in the predictors, in our application the symptom variables. During tree-building it treats 'unknown' as a value like any other, thus allowing those cases for which the partitioning variable is unknown to be passed down to one of the child nodes. This is the only viable approach with our data, since the values of a high proportion of variables are missing for any given case. In classification, evaluation stops if a node is reached for which the variable is unknown. The case is then classified as having the disease most common among the cases at that node.

As with all other methods, CART was tested using the same leave-out-101 strategy. The intention was, for each batch of test cases to grow a decision tree and then prune it using the α-value determined by 10-fold cross-validation. Initially we made available all 169 symptom variables when building the tree. However, the large memory requirement of the S-Plus package meant that it proved impossible to prune these trees using the cross-validation approach, since the machine used (a Sun Sparc-2 workstation with 32Mb of memory) had insufficient random-access memory. This problem was circumvented by restricting the variables available for partitioning to those exhibiting a significant correlation with the disease variable. Correlation was assessed using the χ^2 test on the contingency table defined for each variable, taking 5% as the significance threshold. On average, this reduced the number of symptom variables from 169 to 92.

The error rate obtained with CART was 0.442 (472 errors). This is worse than with iterative partitioning (445 errors), although not significantly so ($p = 0.0814$). We therefore did not pursue recursive partitioning any further for our application. We note that in another study, involving 6387 patients with abdominal pain, recursive partitioning was found to be much less accurate than independence Bayes [Gam91].

10.1.8 Cases with Definitive Diagnoses

Our preliminary conclusion is that knowledge-based programs are not significantly more accurate than purely statistical alternatives. However, the final diagnoses of 375 of the 1270 cases in the database were presumed rather than definite. The possibility remains that the full potential of the knowledge-based programs was not realized because they were trained and tested on cases whose diagnoses were often unreliable. Would the knowledge-based programs outperform the statistical if the diagnostic task were clearer? To answer this question, we repeated the above experiments on the database restricted to the 895 cases whose diagnoses were definite. We adopted the same cross-validation strategy as before, but restricting training and test sets to cases with definite diagnoses. The actual sizes of the training and test sets are shown in Table 10.5: there are a total of 751 amongst the 1068 test cases that have definite diagnoses. The error rates for all programs are shown in Table 10.6. As before, we chose optimal parameters retrospectively. In the case of nearest neighbours, the optimal number of neighbours was 14 for the Hamming metric and 48 for the Bayes metric. Iterative partitioning was optimal with $\alpha = 28.0$. In the case of the neural network, ten hidden units were now found to produce the best results, with the same initial gain of 0.03. When the nearest neighbours classifier with the Bayes metric was trained on 10^6 simulated cases, slightly better results (220 errors) were obtained with $k = 300$ than with $k = 100$ (225 errors).

Table 10.5: Sizes of each training and test set when cases are restricted to those with definite diagnoses.

Test set	Size of test set	Size of training set
B_1	75	815
B_2	62	829
B_3	65	830
B_4	77	818
B_5	60	835
B_6	77	817
B_7	67	826
B_8	70	825
B_9	78	817
B_{10}	75	817
B_{11}	45	849
Total (B)	751	-

Table 10.6: Error rates for all programs on the test set of 751 cases with definite diagnoses, sorted into rank order. See Table 10.5 and the text for details of the cross-validation strategy. The right-hand column lists the corresponding tables in the appendix showing the full discriminant matrices for the five most accurate programs. There is no significant difference in accuracy between those five programs at the 5% significance level using the McNemar test.

Program	Errors	Error rate	Matrix
Neural network (ψ_N, $n = 10$)	202	0.269	Table B.13
Nearest Bayes neighbours (ψ_K, $\delta = \delta_b$, $k = 48$)	205	0.273	Table B.14
Inferential rule-based system (ψ_R, $Q = Q_I$, $T = T_I$)	207	0.276	Table B.15
Independence Bayes (ψ_B, $P = P_I$)	217	0.289	Table B.16
Causal rule-based system (ψ_R, $Q = Q_C$, $T = T_C$)	220	0.293	Table B.17
Small Bayesian network (ψ_B, $P = P_S$)	228	0.304	
Iterative partitioning (ψ_I, $\alpha = 28.0$)	234	0.312	
Large Bayesian network (ψ_B, $P = P_L$)	236	0.314	
Flowchart (ψ_F)	247	0.329	
Exemplar model (ψ_E)	268	0.357	
Nearest Hamming neighbours (ψ_K, $\delta = \delta_I$, $k = 14$)	295	0.393	

There is no statistically significant difference in error rate between the first five programs in Table 10.6. However, it is notable that the inferential rule-based system made fewer errors than the causal rule-based system, whereas previously it made significantly more. This reversal is largely explained by an improvement in the inferential rule-based system's accuracy with respect to abortion (Disease C): see Tables B.15 and B.17. A weakness of the inferential rule-based system would appear to be in confusing more difficult examples of abortion with threatened abortion and with non-specific pain. When the discrimination task is made easier by eliminating the cases in which there is some doubt about the diagnosis, this weakness is less obvious.

10.2 Discussion

In conclusion, therefore, it appears that knowledge-based programs are not significantly more accurate for this application than the best of the purely statistical classifiers. We note that other studies have shown similar results in other applications. An early study by Fox *et al* compared a rule-based system with independence Bayes in the diagnosis of dyspepsia, and found no significant difference [Fox80]. The test set was small, however, consisting only of 50 cases. More recently Ludwig and Heilbronn evaluated a causal network equipped with subjective probability estimates for the diagnosis of chest pain [Lud83]. They found it substantially less accurate than simple logistic regression.

One of the few knowledge-based diagnostic programs to find routine application in medicine is the Pathfinder system due to Heckerman and Nathwani [Hec92c]. The program assists the diagnosis of lymph node disorders by interpreting the features present on histological examination of biopsy specimens. The knowledge representation is a Bayesian network. A more highly connected network was found to be significantly more accurate than a simple conditional independence model, whereas in our application we have found that background knowledge leads to negligible improvement if any. However, firstly, unlike our programs, all probabilities incorporated in Pathfinder are *subjective* estimates provided by one of the authors. De Dombal and colleagues concluded many years ago that independence Bayes classifiers do not work well, especially for rarer conditions, when provided with subjective rather than objective probability estimates [Lea72, Dom78]. Indeed experts have even been shown to be unreliable simply in identifying which symptoms and signs are useful discriminants [Kni85]. Secondly, the accuracy of the two versions of Pathfinder was also assessed subjectively by the same author who provided the original probability estimates. It is perhaps not surprising that the more flexible, highly parameterized model was better able to accommodate the intentions of the expert. Furthermore, an independence Bayes model tends to lead to over-optimistic predictions of posterior probability; the author assessing the 'quality' of the posterior distributions may simply have preferred the more conservative predictions of the more complicated model to those of the independence Bayes model. It is interesting to note that when the two models were compared in their ability to predict the *true* diagnosis from the observations of a non-expert (this is the stated purpose of Pathfinder), no significant difference in accuracy was detected.

In a complementary study to the one described here, we used the causal rule-based system as a simulation model to investigate the limits to diagnostic accuracy achievable by statistical methods [Tod93a]. We concluded that with sufficient training examples (10^5), the independence Bayes classifier is near-optimal: taking interactions into account by various

methods (Lancaster model for pairwise interactions, nearest neighbours using the Bayes metric, neural networks) leads to no significant improvement in diagnostic accuracy. It is therefore not surprising that in the experiments described above, no program has been shown to be significantly more accurate than independence Bayes.

However, one program that has performed as well as independence Bayes is nearest neighbours with the Bayes metric. The nearest neighbours method is particularly suited to medical diagnosis because it is so accountable. The diagnostic prediction of the system regarding a new case is encoded as a small set of previous actual cases. The user is thus able to examine and vet those cases, and decide for himself if they really are representative of the new case. The nearest neighbours method has not been widely adopted because of poor accuracy (e.g. [Ser85]), however our new metric appears to correct this deficiency. Furthermore, in another study [Sta93] we showed that the Bayes metric corresponds as well as Hamming distance to the notion of 'clinical similarity' between case histories. In that study we found it necessary to use the 'anatomical process' variable rather than the 'final diagnosis' variable as a target for computing the posterior distribution so as not to lose information about the side of symptoms and signs. Retrieval of cases with similar Bayesian analyses has also been advocated by de Dombal *et al* [Dom92]. They employ a rather different technique to ours: all cases are retrieved which have the same leading diagnosis as the target case, and whose posterior probabilities lie in the same broad pre-determined interval as that of the target case. This appears wasteful of information about the patient's presentation, and it casts some doubt as to whether the presentations of the retrieved cases are truly 'similar' to the target case as claimed.

Although perhaps not applicable to abdominal pain, the nearest neighbours method does have the additional advantage of taking interactions into account (even if an independence Bayes model is used as the metric), and so it is potentially more accurate than independence Bayes itself. As a final comparison of nearest neighbours and independence Bayes, we measured their accuracies on all 1270 cases in the database (this is permissible since neither program is knowledge-based) using a 'leave-one-out' training strategy (any repeat presentations of the same patient as the test case were also left out from the training set). Figure 10.6 plots error rate against choice of k for nearest neighbours. With the most optimistic value ($k = 29$), the nearest neighbours classifier still only makes five fewer errors than independence Bayes (Table 10.7). This difference is not statistically significant, indeed in as many as 121 of the 1270 cases one of the two programs makes the correct diagnosis while the other does not, so disagreement is considerable. It seems clear from our study that for the diagnosis of acute abdominal pain, there is nothing to be gained by attempting to take interactions into account. Nevertheless, our nearest neighbour program is no less accurate than all the others that we have tried, it is simple and has the additional advantage of accountability. The latter is particularly important in safety-critical fields such as medical diagnosis. Researchers currently using independence Bayes classifiers might like to try this new technique.

Acknowledgements

We are grateful to Tony Hoare for providing helpful suggestions and advice. We would like to thank the Consultants in the Nuffield Department of Obstetrics and Gynaecology, John Radcliffe Hospital, Oxford, for permission to collect and use clinical data from patients under their care, and Paul Macpherson for his invaluable assistance. We are also grateful to

Figure 10.6: Graph of error rate as a function of k for the nearest neighbours method using the Bayes Metric when tested on all 1270 cases in the database using a 'leave-out-one' training strategy. The error rate for independence Bayes (0.374) is also shown for comparison.

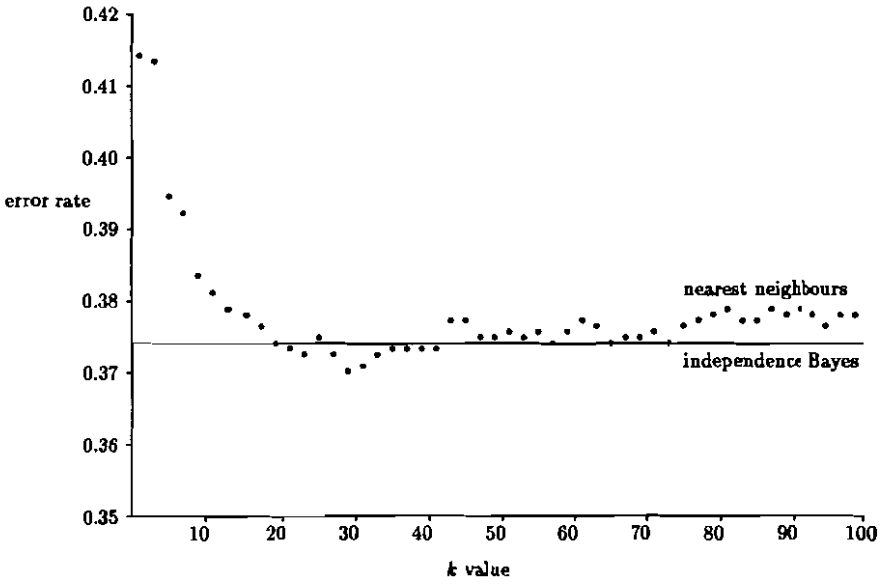


Table 10.7: Error rates for independence Bayes and nearest neighbours using the Bayes metric on all 1270 cases in the database using a 'leave-out-one' training strategy. The right-hand column lists the corresponding tables in the appendix showing the full discriminant matrices.

Program	Errors	Error rate	Matrix
Nearest Bayes neighbours (ψ_K, $\delta = \delta_b$, $k = 29$)	470	0.370	Table B.18
Independence Bayes (ψ_B, $P = P_I$)	475	0.374	Table B.19

Professor F.T. de Dombal for supplying us with some initial cases that were a useful guide in constructing the various knowledge bases.

Richard Stamper is supported by an SERC Research grant (SERC GR/F47077). Bryan Todd is supported by an SERC Advanced Fellowship in Information Technology (SERC B/ITF/269).

Appendix A

Variable Definitions

Definitions are given for all symptom variables and additional variables. Also shown are the possible values of each variable.

A.1 Symptom Variables

A total of 169 symptom variables were recorded.

Personal Details

Age — Age of patient (in decades) at time of presentation.

Characteristics of the Pain

Initial site of pain — Site at which the pain was maximal (or the distribution of the pain if generalized) at onset of the current episode. If the site of pain is not mentioned then it is unknown. However, if the pain is described as being contraction-like, or like a period pain then the site is recorded as 'lower'.

Final site of pain — Site or distribution of the pain at the time the patient presented, or the final site of the pain if the pain had disappeared by the time the patient presented. If no mention is made of the pain changing its distribution then the final site is taken by default to be the same as the initial site.

Duration of pain — Duration (in terms of the order of magnitude) of the present episode of pain. Thus, if the patient has had intermittent pain for five days, then 'days' is recorded. However, if the patient has had constant pain for six hours having been free of pain for three months since a similar previous episode, then 'hours' is entered.

Radiation — Nine Boolean variables independently record any reported radiation of the pain to each of nine sites. If no mention is made of the presence or absence of any radiation, then all nine variables are unrecorded. However, if any mention is made of radiation, then all variables assume a default value of 'false' since it is generally understood that radiation to sites other than those mentioned is absent. Thus for example if the pain is described

simply as radiating to the groin, 'pain radiates to perineum' is recorded as 'true' and the eight other variables are recorded as 'false'.

Aggravating and relieving factors — Ten Boolean variables independently record various factors which are reported as aggravating or relieving the pain. Since these factors are not routinely sought, no default values can be assumed. Thus for example if the patient is described as experiencing period pains on exertion, then the variable 'pain is aggravated by movement' is recorded as 'true'. If the pain is said to be relieved by bed-rest then 'pain is relieved by lying still' is recorded as 'true'.

Severity of pain — By default the severity of the pain is taken to be moderate, since if the pain were mild or severe then the clinician would normally record this in the notes. In particular, the severity is interpreted as if the pain is described as aching or dull, moderate if described as 'not severe' or 'like bad period pains' or 'sharp' or 'pain ++', and severe if described as a 'strong pain'.

Type of pain — Since there is no normal type of pain, no default value can be assumed for this variable. In particular the type is taken to be steady if the pain is described as 'constant' or as a 'dull ache', cramping if described as period-like or a 'severe ache', fluctuating if described as 'constant with exacerbations' or 'stabbing'.

Progress of pain — Course of progression of the pain. By default this is taken to be 'same' because had the pain improved or worsened before presentation, the clinician would normally have recorded this.

Menstrual History

Periods — Usual pattern of periods. For example, if periods were regular until six months ago when they ceased completely then 'regular' is recorded, but if six months ago the periods became irregular then 'irregular' is entered. If no mention is made of the periods being regular or irregular, then if the minimum and maximum cycle lengths have been recorded, the regularity of the periods is calculated from those: the periods are regular precisely when the difference between the minimum and maximum cycle lengths is no greater than five days.

Weeks since reported LMP — If a date for the LMP is available, then the number of days n at presentation since the LMP is calculated. This is converted to a discrete value as follows.

$$\begin{aligned} 0 \leq n < 4 &\quad \text{weeks\_since\_reported\_LMP} = 0 \\ 4 \leq n < 32 &\quad \text{weeks\_since\_reported\_LMP} = 1\_to\_4 \\ 32 \leq n < 46 &\quad \text{weeks\_since\_reported\_LMP} = 5\_to\_6 \\ 46 \leq n &\quad \text{weeks\_since\_reported\_LMP} = 7\_or\_more \end{aligned}$$

If no precise date is given then if possible an estimate is entered instead. This may be the average of a range of possible dates suggested by the patient, or a date calculated retrospectively from the gestational age (if pregnant). A statement such as 'the beginning of March' is interpreted as 1/3/1990. If the date is not mentioned, and a reasonable estimate cannot be made from other information in the notes, then it is unknown.

Time of LMP — Relative timing of the patient's last reported period. For example, if the last period was on time, even if it was six months ago, then 'on time' is recorded. Note

that the timing is *not* recorded as 'late' simply because say 'two months overdue' is written in the notes: this is because the comment refers to some future menstrual period and not to the last known period. No default value can be assumed, but a statement such as 'LMP entirely normal' is taken to mean that the LMP was on time.

Type of LMP — Type of the last reported period. A statement such as 'LMP entirely normal' or an entry such as 'bleeding ++' is taken to mean that the LMP was a moderate flow.

Abnormal Bleeding

Bleeding since LMP — Records whether or not the patient has experienced any abnormal PV bleeding since the last menstrual period, and if so the nature of the bleeding. The reported passage of products takes precedence over the passage of clots. By default if no mention of bleeding is made in the notes then it is assumed by default that none was experienced.

Onset of bleeding — Time elapsed (hours, days, weeks or months) since the onset of any abnormal bleeding following the LMP. If this time is not mentioned explicitly in the notes, and it cannot be deduced from other information, (or if no bleeding has been reported) then it is assumed to be unknown.

Type of bleeding — Severity of any abnormal bleeding since the LMP. If no mention is made of the severity of the bleeding, then (provided that bleeding has been reported) it is assumed by default to be moderate. Specific comments such as 'like a period' or 'bleeding↑' (indicating that the bleeding is worsening) are taken to mean that the bleeding was moderate. However, ' bleeding+++' is taken to mean the bleeding was severe.

Progress of bleeding — Course of progression of any bleeding since the LMP. If no mention is made of the progress of the bleeding, then (provided that bleeding has been reported) it is assumed by default to be 'same'.

Other Symptoms

A total of 29 other variables record the presence of various other symptoms and their time since onset. A symptom is considered to be present if it still constitutes an active clinical problem. Thus if a patient has been troubled by intermenstrual bleeding on and off for years, then it is still considered to be present, even if no bleeding has occurred in the last cycle, provided that this is consistent with the general pattern. Symptoms are assumed by default to be absent, with two exceptions; 'breast symptoms' are not routinely enquired about if the patient is known to be pregnant at presentation, and so are absent by default only if the patient is not known to be pregnant; 'dyspareunia' is not routinely enquired about, and so is unknown unless explicitly mentioned in the notes. Note that pain 'after coitus' is also considered to be internal dyspareunia. If both internal and external dyspareunia are present, then the internal dyspareunia is likely to be much more significant clinically; therefore only a single variable is used to record dyspareunia, the internal variety taking precedence over the external whenever both are present. Note also that reference to the passage of small clots does not in itself constitute menorrhagia, although 'heavy loss' as a description of the usual menstrual flow does. Anorexia, nausea and vomiting are considered to be a progression of the same phenomenon, and are recorded in a single variable. Dizziness and

light-headedness is not distinguished from faintness or actual fainting. Any recent shivering or sweating constitutes a 'recent fever or chill': it is thus implied by the entry 'hot +cold this week' or 'night sweats'. Conversely the single entry '^{O}rigors' implies that there has been no recent fever or chill. The time since onset of any symptom is unknown unless explicitly mentioned.

Obstetric History

Parity — Number of births. If there is evidence in the notes that the clinician has enquired about any miscarriages and/or terminations then by default the parity is assumed to be zero.

Miscarriages — Number of spontaneous abortions. If there is evidence in the notes that the clinician has enquired about the parity and/or terminations then by default it is assumed that there have been no miscarriages.

Terminations — Number of terminations of pregnancy. If there is evidence in the notes that the clinician has enquired about the parity and/or any miscarriages then by default it is assumed that there have been no terminations.

Contraception — Current method of contraception. If the patient is pregnant, then the variable refers to the method of contraception used at the time the patient became pregnant. The entry 'planned pregnancy' indicates that no contraception was used. Similarly, if the patient has been sterilized (previous tubal ligation without subsequent reversal, previous hysterectomy or previous bilateral oophorectomy or salpingectomy) then the current contraception is recorded as 'none'. Also, if the patient has recently been taking fertility drugs, then it is assumed that they are not using contraception.

Infertility — The patient's apparent fertility in the absence of any sterilization procedure. The patient is recorded as being 'infertile' (i.e. infertile or subfertile) if she has been trying to conceive for more than one year, even if she is now pregnant; the infertility is primary if she has never been pregnant, otherwise it is secondary. Reference to 'subfertility' in the notes is interpreted to mean 'infertility'. If no mention of infertility or subfertility is made in the notes, then it is assumed that the patient does not have infertility, unless the patient has had some surgical procedure which renders her sterile.

Pregnancy — Records the apparent possibility of pregnancy on presentation. If the patient has had a surgical procednre which has rendered her sterile, then pregnancy is recorded as impossible. Otherwise, however, if no mention is made of the impossibility or certainty of pregnancy, then it is unknown

Past Medical History

Past History — A total of 12 variables record previous significant medical conditions. By default it is assumed that none have occurred, except for 'complete abortion': it cannot be assumed by default that there have been no complete abortions if miscarriages have been reported, unless the patient has also had an ERPC in the past. The time since the last complete abortion is also recorded in one of the 12 variables. (Note that the entry '^{O}illnesses' conveys no information about any of these conditions, and is ignored.)

Previous surgery — A total of 34 variables record previous surgical procedures. By default it is assumed that none have been performed, except for 'ERPC': it cannot be assumed by default that the patient has never had an ERPC if miscarriages have been reported. One of the 34 variables ('previous other uterine instrumentation') is intended to record all procedures involving uterine instrumentation other than D+C, ERPC and TOP: for example, insertion of IUCD, cervical catheterization, and hysterosalpingogram investigation. Another variable ('previous laparotomy') records all intraperitoneal surgery other than the procedures explicitly recorded by the other 11 variables: for example, cholecystectomy.

Drug History

Recent medication — Three Boolean variables independently record whether antibiotics, analgesia and/or fertility drugs have been administered recently. 'Recent antibiotics' are those which were started at least 24 hours before presentation, and have not been stopped earlier than 24 hours before presentation. 'Recent analgesia' refers to strong analgesia (e.g. pethidine, morphine, or temgesic) only, and administered in the 24 hours preceding presentation. 'Recent fertility drugs' are those administered within one year of presentation. By default it is assumed that none were administered.

Cigarettes per day — Average number of cigarettes normally smoked per day. If a range is given (e.g. 15—20) then the average is taken and rounded to a whole number (18). If the patient stopped smoking one month or more ago then the patient is recorded as a being a non-smoker, if less than one month, then the usual number of cigarettes is recorded. The average number per day n is assigned to one of three categories as follows.

$$\begin{aligned} n = 0 &\quad \text{cigarettes\_per\_day} = \text{none} \\ 1 \leq n < 10 &\quad \text{cigarettes\_per\_day} = \text{less\_than\_10} \\ 10 \leq n &\quad \text{cigarettes\_per\_day} = \text{10\_or\_more} \end{aligned}$$

Alcohol consumption — Average amount of alcohol normally consumed. If not mentioned in the notes then it is unknown. Notice that no distinction is made between occasional consumption and moderate consumption.

General Examination

Overweight — Records whether the patient appears obese. If no mention is made of the presence or absence of obesity, then it is unknown unless the patient has been weighed. In the latter case, the patient is overweight precisely when they weigh 70kg or more.

Pulse — Records whether or not the patient has a tachycardia (pulse rate of 100 per minute or more) on presentation to the admitting team. If more than one figure is available, the preferred figure is that in the clerking notes, in the nursing notes, on the bed chart or in the GP's referral letter (decreasing order of preference). If no mention is made of the pulse rate then it is unknown.

Mean BP — Records whether or not the patient is hypotensive (mean BP strictly less than 70mmHg) on presentation to the admitting team. If more than one figure is available, the preferred figure is that in the clerking notes, in the nursing notes, on the bed chart or

in the GP's referral letter (decreasing order of preference). If no mention is made of the pulse rate then it is unknown.

Temperature — Records the patient's temperature at presentation to the admitting team. If more than one figure is available, the preferred figure is that in the clerking notes, in the nursing notes, on the bed chart or in the GP's referral letter (decreasing order of preference). If no mention is made of the pulse rate then it is unknown.

Colour — Patient's facial colour. An entry such as 'pallor' indicates that the colour is normal. An entry such as 'pale conjunctivae and tongue' indicates that the patient is pale. Colour is assumed by default to be normal unless there is a comment in the notes to the effect that the patient is toxic or shocked.

Mood — Patient's mood. An entry such as 'not distressed' or 'well' indicates a normal mood, 'tearful' indicates that the patient is distressed, and 'shaken' indicates that the patient is psychologically shocked.

Sweating — Records whether the patient is sweating. Sweating is absent by default unless there is a comment in the notes to the effect that the patient is toxic or shocked.

Dehydrated — Records whether the patient is clinically dehydrated. A comment such as 'fit and well' implies that the patient is not clinically dehydrated. In any case, by default the patient is assumed not to be clinically dehydrated, because the clinician would record this.

Abdominal Examination

Site of tenderness — Site at which abdominal tenderness is maximal (or the distribution of the tenderness if it is generalized). This is usually deduced from diagrams drawn in the notes. Sometimes the comment 'tender fundus' is found; this indicates that the site of tenderness was the lower abdomen. If tenderness is not mentioned then it is assumed by default that none was found provided that it appears from the notes that the abdomen was examined.

Other abdominal signs — Six Boolean variables independently describe the other findings on abdominal examination. All of these are assumed to be false (absent) by default provided that it appears from the notes that the abdomen was examined. Note that a mass which has been clearly identified as an enlarged organ (for example, liver or uterus) is not included as a 'mass'. Also, no distinction is made between a suspected mass and an unequivocal mass.

Bowel sounds — Nature of the bowel sounds. An entry 'BS present' or 'BS$\surd$' indicates that the bowel sounds are normal, and 'BS active' indicates that they are increased. If no mention is made of the bowel sounds then they are unknown because auscultation for bowel sounds is not a routine part of the examination.

Vaginal Examination

PV tenderness — Three variables record any tenderness in the right adnexa, in the left adnexa, and centrally, respectively. If general tenderness is noted, then this indicates tenderness in all three areas. (Pain on speculum examination is interpreted as general tenderness in the absence of more specific information.) Central tenderness includes both uterine tenderness and tenderness in the Pouch of Douglas. By default it is assumed that

no tenderness is present, provided that it appears from the notes that a PV examination, however limited, was performed.

PV mass — Two variables record whether a mass was detected, or even suspected, in each adnexa. By default it is assumed that no mass was detected provided that it appears from the notes that the adnexae were examined.

Cervical excitation — Records whether cervical excitation is present or not. It is assumed by default that none is present provided that it appears from the notes that the adnexae were examined (in which case an attempt to elicit cervical excitation would have been made routinely), and the patient had not had a hysterectomy. Note that 'suspected excitation' or 'mild excitation' is regarded as actual excitation.

Cervix — State of the cervical os. Note that no distinction is made between 'half open' and 'open'. The cervix is assumed by default to be closed provided that the adnexae have been examined, unless the patient has had a hysterectomy.

Uterus enlarged — Records whether or not the uterus is enlarged. However, it is not always possible to assess the size of the uterus when significant tenderness is present. Nevertheless, if the size were assessed, then it would nearly always be recorded, even if it were normal. Therefore it is not safe to assume by default that the uterus is of normal size.

Size for dates — Records the relative size for dates of the uterus. A uterus is small for dates precisely when it is three or more weeks smaller than expected for dates (calculated to the nearest whole week). Similarly a uterus is large for dates precisely when it is three or more weeks larger than expected for dates. No default assumption about the size for dates can be made.

Uterus — Records whether or not the uterus is anteverted. However, it is not always possible to assess the position of the uterus when significant tenderness is present. Nevertheless, if the position were assessed, then it would nearly always be recorded, even if it were anteverted. Therefore it is not safe to assume by default that the uterus is anteverted.

Speculum — Four variables record findings on speculum examination. If findings are not mentioned in the notes, then the findings on EUA are recorded instead if EUA was performed within 24 hours from presentation. A discharge is classed as purulent unless it is described as 'slight', 'clear' or 'normal', in which cases it is classed as clear. The variable 'speculum blood' records whether blood and/or products of conception were seen, the latter subsuming the former in importance. Speculum examination is not always performed, however if it appears from the notes that it was performed then all findings are assumed by default to be negative.

Blood and Urine Tests

Pregnancy test — Sensitivity and result of the pregnancy test. For example, the entry 'Prognosticon ?+ve' indicates that the pregnancy test is equivocal; 'Ramp +ve' indicates that the pregnancy test is 'positive high'. If both a low and a high sensitivity test have been performed, then the more significant test is the relevant one. In each case this means the high sensitivity test, unless the results of both tests are the same, in which case the low sensitivity test is the more significant. No default assumption can be made about the result of the pregnancy test.

Haemoglobin — Records whether the haemoglobin level (g/dl) is low (Hb $<$ 10.0), normal (10.0 $\leq$ Hb $<$ 14.0), or high (14.0 $\leq$ Hb). No default assumption can be made.

white_cell_count — Records whether the white cell count (cells per nanolitre) is low (WCC $<$ 4.0), normal (4.0 $\leq$ WCC $<$ 11.0), or high (11.0 $\leq$ WCC). No default assumption can be made.

Platelets — Records whether the platelet count (platelets per nanolitre) is low (PC $<$ 140), normal (140 $\leq$ PC $<$ 440), or high (440 $\leq$ PC). No default assumption can be made.

erythrocyte_sedimentation_rate — Records whether or not the ESR is elevated (12 or more mm/hr). No default value can be assumed.

Urinalysis — Four variables independently record the findings on urinalysis. The distinction between '+' and '++' is dropped. Thus the levels are translated as follows.

nil	$\equiv$	none
+	$\equiv$	minimal
++	$\equiv$	moderate
+++	$\equiv$	moderate

No default values can be assumed.

Urine microscopy — Three variables record the results of urine microscopy. The number n of cells per field is translated as follows.

$n = 0$	$\equiv$	none
$1 \leq n < 19$	$\equiv$	minimal
$20 \leq n$	$\equiv$	moderate

A reference to 'occasional cells' is interpreted as 'minimal'. No default values can be assumed.

Ultrasound Examination

The results of ultrasound examination are relevant if the investigation was carried out promptly *after* the patient was clerked. If the investigation was performed *before* the patient was clerked, or more than 24 hours later, then the results are admissible only if, in the light of the final diagnosis, one would not normally have expected them to be any different.

Ultrasound Type — Type of ultrasound examination performed. It is assumed to be abdominal unless otherwise stated.

Ultrasound adnexae — Two variables independently record the ultrasound findings in the left and right adnexae, respectively. Note that small echosonic areas surrounded by thickened areas are interpreted as cysts. The adnexae are assumed to be normal by default.

Ultrasound Pouch of Douglas — Records whether or not fluid was detected in the Pouch of Douglas. It is assumed by default that none was present.

Ultrasound uterine wall — Records whether fibroids were detected in the uterine wall. It is assumed by default that none were found.

Ultrasound uterine cavity — Contents of the uterus. If reference is made to a gestational sac being seen that is too small for dates then this is entered as a missed abortion rather

than as a gestational sac. An entry such as 'intrauterine pregnancy ≈ 8/40' is interpreted as 'foetal pole'. By default it is assumed that an empty uterine cavity was found.

Ultrasound uterine pregnancies — Records whether or not more than one uterine pregnancy was detected in the case that the uterus is non-empty. If the uterus is empty then this variable is left unrecorded. Otherwise it is assumed by default that only a single pregnancy was detected, provided that a gestational sac, foetal pole, foetal heart (with or without an associated haematoma), missed abortion, or hydatidiform mole was seen in the uterine cavity.

A.2 Additional Variables

A total of 53 additional variables were recorded.

Menstrual Periods

Weeks since actual LMP — Weeks since the start of the patient's *actual* LMP. This may well be different from the reported one. Unless there is good reason to doubt the accuracy of the reported LMP, the time is calculated from the patient's dates. However, when other evidence as to the patient's stage of gestation is taken into account, it may appear more likely that the reported LMP was in fact an episode of abnormal bleeding. Similarly, if the patient presents with bleeding of a day or two's duration and non-specific pain, it may appear more probable that the current episode of bleeding is menstrual and, in fact, represents the actual LMP. However, if the most recent episode of bleeding has occurred at a time, or for a duration, that is not typical of the patient's menstrual pattern, then it should not be designated the actual LMP. If the number of weeks since the reported LMP is unknown, then so too is the number of weeks since the actual LMP.

Menstruating — Records whether the patient is menstruating at the time of presentation.

Pregnancy

Pregnancy since LMP — Records whether the patient has become pregnant since the *actual* LMP, and if so, the site of the pregnancy. If a patient has both an ectopic and an intrauterine pregnancy, then the ectopic site is recorded preferentially.

Raised HCG — Records the time for which the HCG level has been significantly elevated. If the HCG level was previously elevated, but has since fallen to normal, then the value 'false' is entered. Since HCG levels are not generally repeatedly measured in the way that would be required here, the decision as to the value of this variable must be based on pregnancy tests that have been performed, the duration of the patient's symptoms of pregnancy, and the gestational age of the pregnancy.

Raised progesterone — Records the time for which the progesterone level has been significantly elevated. If the progesterone level was previously elevated, but has since fallen to normal, then the value 'false' is entered.

Table A.1: Possible values for variables 1 to 60

No.	Name	Values
1	age	in_teens, in_20s, in_30s, in_40s
2	initial_site_of_pain	RUQ, LUQ, RLQ, LLQ, central, upper, lower, right, left, generalised, right_loin, left_loin, back, other
3	final_site_of_pain	RUQ, LUQ, RLQ, LLQ, central, upper, lower, right, left, generalised, right_loin, left_loin, back, other
4	duration_of_pain	hours, days, weeks, months, years
5	pain_radiates_to_right_shoulder	false, true
6	pain_radiates_to_left_shoulder	false, true
7	pain_radiates_to_right_loin	false, true
8	pain_radiates_to_left_loin	false, true
9	pain_radiates_to_back	false, true
10	pain_radiates_suprapubically	false, true
11	pain_radiates_to_perineum	false, true
12	pain_radiates_to_right_leg	false, true
13	pain_radiates_to_left_leg	false, true
14	pain_is_aggravated_by_lying_flat	false, true
15	pain_is_aggravated_by_movement	false, true
16	pain_is_aggravated_by_retching	false, true
17	pain_is_aggravated_by_coughing	false, true
18	pain_is_aggravated_by_respiration	false, true
19	pain_is_aggravated_by_food	false, true
20	pain_is_relieved_by_lying_still	false, true
21	pain_is_relieved_by_vomiting	false, true
22	pain_is_relieved_by_antacids	false, true
23	pain_is_relieved_by_food	false, true
24	severity_of_pain	mild, moderate, severe
25	type_of_pain	intermittent, steady, colicky, cramping, fluctuating, other
26	progress_of_pain	stopped, better, same, worse
27	periods	regular, irregular, none_yet
28	weeks_since_reported_LMP	0, 1_to_4, 5_to_6, 7_or_more
29	time_of_LMP	early, on_time, late
30	type_of_LMP	light, moderate, heavy
31	bleeding_since_LMP	false, blood, clots, products
32	onset_of_bleeding	hours, days, weeks, months, years
33	type_of_bleeding	light, moderate, severe
34	progress_of_bleeding	stopped, better, same, worse
35	recent_fever_or_chill	false, true
36	anorexia_nausea_or_vomiting	false, anorexia, nausea, vomiting
37	duration_of_anorexia_nausea_vomiting	hours, days, weeks, months, years
38	constipation	false, true
39	duration_of_constipation	hours, days, weeks, months, years
40	diarrhoea	false, true
41	duration_of_diarrhoea	hours, days, weeks, months, years
42	frequency	false, true
43	duration_of_frequency	hours, days, weeks, months, years
44	dysuria	false, true
45	duration_of_dysuria	hours, days, weeks, months, years
46	haematuria	false, true
47	duration_of_haematuria	hours, days, weeks, months, years
48	discharge	false, true
49	duration_of_discharge	hours, days, weeks, months, years
50	breast_tenderness_or_enlargement	false, true
51	duration_of_breast_symptoms	hours, days, weeks, months, years
52	faintness	false, true
53	duration_of_faintness	hours, days, weeks, months, years
54	dysmenorrhoea	false, true
55	duration_of_dysmenorrhoea	hours, days, weeks, months, years
56	menorrhagia	false, true
57	duration_of_menorrhagia	hours, days, weeks, months, years
58	postcoital_bleeding	false, true
59	duration_of_postcoital_bleeding	hours, days, weeks, months, years
60	intermenstrual_bleeding	false, true

Table A.2: Possible values for variables 61 to 120

No.	Name	Values
61	duration_of_intermenstrual_bleeding	hours, days, weeks, months, years
62	dyspareunia	false, external, internal
63	duration_of_dyspareunia	hours, days, weeks, months, years
64	parity	0, 1, >1
65	miscarriages	0, 1, >1
66	terminations	0, 1, >1
67	contraception	pill, depo_provera, IUCD, condom, cap, diaphragm, vasectomy, none, other
68	infertility	primary, secondary, none
69	pregnancy	impossible, established, indeterminate
70	past_history_of_PID	false, true
71	past_history_of_ovarian_cyst	false, true
72	past_history_of_ectopic_pregnancy	false, true
73	past_history_of_endometriosis	false, true
74	past_history_of_fibroids	false, true
75	past_history_of_threatened_abortion	false, true
76	past_history_of_complete_abortion	false, true
77	time_since_last_complete_abortion	hours, days, weeks, months, years
78	past_history_of_UTI	false, true
79	past_history_of_renal_calculus	false, true
80	past_history_of_irritable_colon	false, true
81	similar_previous_episode	false, true
82	previous_appendicectomy	false, true
83	time_since_appendicectomy	hours, days, weeks, months, years
84	previous_laparoscopy	false, true
85	time_since_laparoscopy	hours, days, weeks, months, years
86	previous_laparotomy	false, true
87	time_since_laparotomy	hours, days, weeks, months, years
88	previous_cervical_surgery	false, true
89	time_since_cervical_surgery	hours, days, weeks, months, years
90	previous_other_uterine_instrumentation	false, true
91	time_since_other_uterine_instrumentation	hours, days, weeks, months, years
92	previous_tubal_ligation	false, true
93	time_since_tubal_ligation	hours, days, weeks, months, years
94	previous_conservative_tubal_surgery	false, true
95	time_since_conservative_tubal_surgery	hours, days, weeks, months, years
96	previous_reversal_of_sterilization	false, true
97	time_since_reversal_of_sterilization	hours, days, weeks, months, years
98	previous_right_oophorectomy	false, true
99	time_since_right_oophorectomy	hours, days, weeks, months, years
100	previous_left_oophorectomy	false, true
101	time_since_left_oophorectomy	hours, days, weeks, months, years
102	previous_right_salpingectomy	false, true
103	time_since_right_salpingectomy	hours, days, weeks, months, years
104	previous_left_salpingectomy	false, true
105	time_since_left_salpingectomy	hours, days, weeks, months, years
106	previous_Caesarian_section	false, true
107	time_since_Caesarian_section	hours, days, weeks, months, years
108	previous_hysterectomy	false, true
109	time_since_hysterectomy	hours, days, weeks, months, years
110	previous_termination	false, true
111	time_since_termination	hours, days, weeks, months, years
112	previous_D+C	false, true
113	time_since_D+C	hours, days, weeks, months, years
114	previous_ERPC	false, true
115	time_since_ERPC	hours, days, weeks, months, years
116	recent_antibiotics	false, true
117	recent_analgesia	false, true
118	recent_fertility_drugs	false, true
119	cigarettes_per_day	none, less_than_10, 10_or_more
120	alcohol_consumption	none, moderate, heavy

Table A.3: Possible values for variables 121 to 169

No.	Name	Values
121	overweight	false, true
122	pulse	less_than_100, more_than_100
123	meanBP	less_than_70mmHg, more_than_70mmHg
124	temperature	less_than_37.5, 37.5_to_38.0, 38.0_or_more
125	colour	normal, pale, flushed, other
126	mood	normal, distressed, shocked, miserable, anxious, tired, other
127	sweating	false, true
128	clinically_dehydrated	false, true
129	site_of_tenderness	RUQ, LUQ, RLQ, LLQ, central, upper, lower, right, left, generalised, right_loin, left_loin, back, none
130	abdominal_guarding	false, true
131	abdominal_rebound	false, true
132	abdominal_rigidity	false, true
133	abdominal_distension	false, true
134	abdominal_mass	false, true
135	ascites	false, true
136	bowel_sounds	absent, decreased, normal, increased
137	PV_tenderness_to_right	false, true
138	PV_tenderness_to_left	false, true
139	PV_tenderness_centrally	false, true
140	PV_mass_to_right	false, suspected, true
141	PV_mass_to_left	false, suspected, true
142	cervical_excitation	false, true
143	cervix	closed, open
144	uterus_enlarged	false, true
145	size_for_dates	small, normal, large
146	uterus	anteverted, retroverted
147	speculum_discharge	none, purulent, brown, white, clear
148	speculum_blood	false, blood, products
149	speculum_cervical_erosion	false, true
150	speculum_vaginal_wall_cyst	false, true, ruptured
151	pregnancy_test	negative_high, negative_low, equivocal, positive_high, positive_low
152	haemoglobin	less_than_10.0, 10.0_to_14.0, 14.0_or_more
153	white_cell_count	less_than_4.0, 4.0_to_11.0, 11.0_or_more
154	platelets	less_than_140, 140_to_440, 440_or_more
155	erythrocyte_sedimentation_rate	less_than_12, 12_or_more
156	urinalysis_protein	none, minimal, moderate
157	urinalysis_glucose	none, minimal, moderate
158	urinalysis_ketones	none, minimal, moderate
159	urinalysis_blood	none, minimal, moderate
160	urine_microscopy_pus_cells	none, minimal, moderate
161	urine_microscopy_red_cells	none, minimal, moderate
162	urine_microscopy_squames	none, minimal, moderate
163	ultrasound_type	abdominal, vaginal
164	ultrasound_right_adnexa	normal, enlarged, mass, cyst
165	ultrasound_left_adnexa	normal, enlarged, mass, cyst
166	ultrasound_Pouch_of_Douglas	fluid, no_fluid
167	ultrasound_uterine_wall	normal, fibroids
168	ultrasound_uterine_cavity	empty, thickened_endometrium, gestational_sac, foetal_pole, foetal_heart, foetal_heart+haematoma, retained_products, missed_abortion, mole
169	ultrasound_uterine_pregnancies	single, multiple

Abortion

The stage of the abortion process is described by five variables. In each case, the process of abortion refers only to the *last* pregnancy, and only if the process began within the last 12 months.

Threatened abortion — This has value 'false' unless currently present, in which case the time since onset is recorded.

Inevitable abortion — This has value 'false' unless currently present (i.e. at presentation has not proceeded to incomplete or complete abortion), in which case the time since onset is recorded. If a patient has abdominal pain due to a missed abortion, then abortion is held to be inevitable rather than threatened. The time since onset of pain or bleeding is recorded.

Incomplete abortion — This has value 'false' unless currently present (i.e. ERPC has not yet been performed), in which case the time since onset is recorded.

Complete abortion — The time that it happened is recorded.

Missed abortion — This has value 'true' if there is evidence that the foetus died *in utero*, otherwise 'false'.

Uterine State

Uterine contractions — Records whether uterine contractions sufficient to cause the presenting episode of abdominal pain have occurred and, if so, the time since their onset. If the time since onset of the abdominal pain has not been recorded, then the time since onset of the uterine contractions must be inferred from time since onset of bleeding if known (the bleeding usually precedes the pain during abortion), and the gestational age of the pregnancy. Typically the contractions will have started days prior to presentation, and this is taken as a default value in the absence of any other clue.

Progress of uterine contractions — Records the progress of uterine contractions if they have occurred. A good guide is the progress of the pain. If that variable has not been recorded, then a default value of 'same' is assumed.

Strength of uterine contractions — Records the strength of uterine contractions if they have occurred. A good guide is the severity of the pain. If that variable has not been recorded, then a default value of 'moderate' is assumed.

Uterine bleeding — Records whether abnormal uterine PV bleeding has occurred since the *actual* LMP and, if so, the time since onset. If the patient is menstruating then this variable necessarily has value 'false'. Unless confusion has occurred over the identification of the LMP, the value of this variable is determined by the time since onset of abnormal bleeding reported by the patient. The exception being that the bleeding is minimal and has not been noticed by the patient: such bleeding may explain the presence of blood contaminating the urine, or evidence of uterine bleeding may be seen on speculum examination. Note that an episode of bleeding that is attributed to implantation haemorrhage is not regarded as abnormal bleeding.

Progress of uterine bleeding — Records the apparent progress of the actual uterine bleeding. Unless there is evidence to the contrary, it is assumed to have value 'same'.

Severity of uterine bleeding — Records the apparent severity of the actual uterine bleeding. Unless there is evidence to the contrary, it is assumed to have value 'moderate'. If the patient reports moderate or heavy bleeding with clots, then this indicates that the severity is 'heavy'. If clots were reported, but the bleeding was said to be only light, then the value is taken to be 'moderate'. In the absence of clots, the value is taken to be that indicated by the patient. The value 'minimal' is reserved for bleeding that is so slight that it has gone unnoticed by the patient.

Implantation haemorrhage since LMP — Records whether or not an implantation haemorrhage has occurred since the actual LMP.

Specific Conditions

Retained products — Records the time for which products have been retained following an ERPC, TOP, or delivery. The variable has value 'false' if the retained products have been previously removed (by another ERPC).

Ectopic pregnancy — Two variables record whether an ectopic pregnancy is present on the left or right side, respectively, and whether it has ruptured at the time of presentation. If so, the value indicates whether the rupture was into the peritoneal cavity or the mesosalpinx.

Hyperemesis gravidarum — Records whether the patient has hyperemesis gravidarum at the time of presentation, and if so, the time since its onset.

Abdominal wall strain — Records whether the patient has an abdominal wall strained sufficiently (typically by vomiting) to cause abdominal pain. If so, the time since onset is recorded.

Ovarian cyst — Two variables record whether a cyst of at least 2cm diameter is present on the left and right ovaries, respectively. If so, the cyst is classified as follows:

- 'asymptomatic' - the cyst is not causing pain, and is uncomplicated.
- 'symptomatic' - the cyst is causing pain in the absence of an identifiable complication.
- 'haemorrhagic' - the complication is principally haemorrhage *into* the cyst (with possibly some leakage into the peritoneal cavity).
- 'ruptured' - the complication is principally rupture of the cyst into the peritoneal cavity (with possibly a significant haemorrhage).
- 'torted' - the cyst has torted (and possibly suffered one of the above complications secondarily).

Salpingitis — Two variables record the presence of left and/or right *symptomatic* salpingitis respectively, and if so, the estimated time since onset of the infection. A hydrosalpinx must be inflamed if it is causing pain, and similarly, the presence of a pyosalpinx obviously implies salpingitis. Although salpingitis is usually bilateral, if the symptoms and signs are clearly localized to one or other side, then salpingitis if recorded as present only on that side.

Chronic PID — Records whether the patient has any of the chronic sequelae of PID (adhesions, hydrosalpinges, infertility). It does not denote active infection.

Hydrosalpinx — Two variables record whether the patient has left and/or right hydrosalpinges at the time of presentation. If, for example, the patient has a left hydrosalpinx that has become infected and formed a pyosalpinx, then the patient is regarded as having a left hydrosalpinx (as well as a pyosalpinx).

Pyosalpinx — Two variables record whether left and/or right pyosalpinges are present at the time of presentation (irrespective of whether they have ruptured). Two other variables record whether either pyosalpinx is ruptured at the time of presentation.

Pyelonephritis — Two variables record whether the patient has an acute left and/or right pyelonephritis at the time of presentation, and if so, the time since onset of the infection.

Acute cystitis — Records whether the patient has cystitis at the time of presentation, and if so, the time since onset of the infection. Notice that acute cystitis and acute pyelonephritis are not mutually exclusive, and often coexist.

Ureteric colic — Two variables record whether the patient has a left or right ureteric colic at the time of presentation. If so, the time since onset is recorded. Although the two conditions rarely coexist, they are not mutually exclusive.

Microscopic haematuria — Records whether the patient has microscopic haematuria (not attributable to contamination through PV bleeding) at the time of presentation.

Microscopic pyuria — Records whether the patient has microscopic pyuria (not attributable to contamination through PV discharge) at the time of presentation.

Fibroids — Records whether the patient has fibroids at the time of presentation.

Acute red degeneration — Records whether fibroids are undergoing acute degeneration.

Endometriosis — Records whether the patient has endometriosis. If so, the endometriosis is classified as 'asymptomatic' if it is not causing pain or tenderness, otherwise 'RLQ' or 'LLQ' or 'bilateral' according to its site if symptomatic. Note that bilateral pelvic endometriosis is recorded as 'RLQ' if the symptoms and signs are clearly localized to the RIF (for example).

Peritoneal cavity — Records the contents of the peritoneal cavity at the time of presentation. The following rules help determine the appropriate value.

- 'empty' - the peritoneal cavity is empty, or contains only normal peritoneal fluid.
- 'ascitic fluid' - ascites is present.
- 'free pus' - Pus is free in the peritoneal cavity (from a ruptured pyosalpinx or appendix, for example).
- 'cystic fluid' - The contents of an ovarian cyst have ruptured into the peritoneal cavity (without any bleeding).
- 'minimal haemoperitoneum' - a small quantity (e.g. 50mls or less) of bloodstained fluid is present in the peritoneal cavity.
- 'moderate haemoperitoneum' - frank blood is present in the peritoneal cavity, but not in a quantity sufficient to cause hypovolaemia.
- 'massive haemoperitoneum' - profuse active bleeding is taking place sufficient to risk hypovolaemia.

Peritoneal irritation — records whether the contents of the peritoneal cavity are causing any evidence of irritation (pain, tenderness, rebound, cervical excitation, ileus, diarrhoea, freqnency) at the time of presentation.

Ileus — Records whether the patient had a state of ileus at presentation.

Pelvic collection — Records whether any collection that might be palpable is present in the pelvis at the time of presentation, and if so its nature. For example, the entry 'Blood clot ++in POD' in the operation notes indicates that 'haematoma' should be recorded as the value of the variable 'pelvic collection'.

Cervical erosion — Records whether a cervical erosion is present.

Irritable colon — Records whether the patient has symptoms of irritable bowel syndrome at presentation.

Adhesions — Records whether the patient has pelvic or abdominal adhesions of any kind.

Hyperstimulation — Records whether the patient is being hyperstimulated. No distinction is made between therapeutic stimulation, and that caused by a hydatidiform mole.

Acute appendicitis — Records whether the patient has acute appendicitis at presentation.

Abdominal wall haematoma — Records whether the patient has a haematoma in the abdominal wall at presentation (e.g. due to spontaneous rupture of an inferior epigastric artery, or secondary to laparoscopy).

Table A.4: Possible values for additional variables 170 to 222.

No.	Name	Values
170	weeks_since_actual_LMP	0, 1 to 4, 5 to 8, 7_or_more
171	menstruating	false, true
172	pregnancy_since_LMP	false, uterine, left_tubal, right_tubal, left_ovarian, right_ovarian, hydatidiform_mole
173	raised_HCG	false, days, weeks, months
174	raised_progesterone	false, days, weeks, months
175	threatened_abortion	false, hours, days, weeks, months
176	inevitable_abortion	false, hours, days, weeks, months
177	incomplete_abortion	false, hours, days, weeks, months
178	complete_abortion	false, hours, days, weeks, months
179	missed_abortion	false, true
180	uterine_contractions	false, hours, days, weeks, months
181	progress_of_uterine_contractions	stopped, better, same, worse
182	strength_of_uterine_contractions	mild, moderate, severe
183	uterine_bleeding	false, hours, days, weeks, months
184	progress_of_uterine_bleeding	stopped, better, same, worse
185	severity_of_uterine_bleeding	minimal, light, moderate, heavy
186	implantation_haemorrhage_since_LMP	false, true
187	retained_products	false, days, weeks, months
188	left_ectopic_pregnancy	false, unruptured, ruptured_into_mesosalpinx,
189	right_ectopic_pregnancy	false, unruptured, ruptured_into_mesosalpinx, ruptured_into_peritoneal_cavity
190	hyperemesis_gravidarum	false, hours, days, weeks, months
191	abdominal_wall_strain	false, hours, days, weeks, months
192	left_ovarian_cyst	false, asymptomatic, symptomatic, haemorrhagic, ruptured, torted
193	right_ovarian_cyst	false, asymptomatic, symptomatic, haemorrhagic, ruptured, torted
194	left_salpingitis	false, hours, days, weeks, months
195	right_salpingitis	false, hours, days, weeks, months
196	chronic_PID	false, true
197	left_hydrosalpinx	false, true
198	right_hydrosalpinx	false, true
199	left_pyosalpinx	false, true
200	right_pyosalpinx	false, true
201	ruptured_left_pyosalpinx	false, true
202	ruptured_right_pyosalpinx	false, true
203	acute_left_pyelonephritis	false, hours, days, weeks, months
204	acute_right_pyelonephritis	false, hours, days, weeks, months
205	acute_cystitis	false, hours, days, weeks, months
206	left_ureteric_colic	false, hours, days, weeks, months
207	right_ureteric_colic	false, hours, days, weeks, months
208	microscopic_haematuria	false, true
209	microscopic_pyuria	false, true
210	fibroids	false, true
211	acute_red_degeneration	false, true
212	endometriosis	false, asymptomatic, RLQ, LLQ, bilateral
213	peritoneal_cavity	empty, ascitic_fluid, free_pus, cystic_fluid, minimal_haemoperitoneum, moderate_haemoperitoneum, massive_haemoperitoneum
214	peritoneal_irritation	false, true
215	ileus	false, true
216	pelvic_collection	false, haematoma, infected_haematoma, abscess
217	cervical_erosion	false, true
218	irritable_colon	false, true
219	adhesions	false, true
220	hyperstimulation	false, true
221	acute_appendicitis	false, true
222	abdominal_wall_haematoma	false, true

Appendix B

Discrimination Matrices

All the discriminant matrices are shown. See Table 8.1 on Page 51 for the key to the diagnoses.

Table B.1: Discriminant matrix for nearest neighbours with the Bayes metric (ψ_K, $\delta = \delta_b$, $k = 100$) when tested on all 1068 test cases and trained on 10^6 cases generated from the causal rule-based system (Q_C, T_C), using a 'leave-out-101' training strategy. Error rate = 0.352.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	113	1	26	4	-	4	40	4	10	-	1	-	1	1	-	-	2	-	2	209
	B	7	44	15	-	-	6	2	-	5	-	-	-	-	-	-	-	-	-	-	79
	C	9	6	378	1	3	8	3	-	-	-	-	-	-	-	-	-	1	-	1	410
	D	3	-	10	23	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	6	-	9	1	-	28	11	-	1	-	-	-	-	-	-	-	-	-	-	56
	G	26	-	7	7	-	1	75	4	8	-	-	-	-	-	-	-	-	-	1	129
	H	11	-	1	-	1	3	5	3	5	-	-	-	-	-	-	-	-	-	-	29
	I	13	-	-	-	-	2	12	-	13	-	-	-	-	-	-	-	1	-	-	41
	J	1	-	-	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	3
	K	2	1	-	-	-	-	1	-	-	-	3	-	-	-	-	-	-	-	-	7
	L	-	-	-	-	-	-	-	-	2	-	-	0	-	-	-	-	-	-	-	2
	M	3	-	1	-	-	-	4	-	1	-	-	-	0	-	-	-	-	-	-	9
	N	6	-	1	-	-	-	6	1	1	-	1	-	-	2	-	-	-	-	-	18
	O	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	1
	P	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	9	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	6	-	-	2	-	2	8	2	1	-	1	-	-	-	-	-	-	-	0	22
		209	52	450	38	4	54	171	15	47	1	6	0	1	3	0	0	13	0	4	1068

Table B.2: Discriminant matrix for nearest neighbours using the Bayes metric (ψ_K, $\delta = \delta_b$, $k = 19$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.362.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	136	9	17	4	-	4	22	6	8	-	-	-	1	-	-	-	-	-	2	209
	B	12	47	13	1	-	3	2	1	-	-	-	-	-	-	-	-	-	-	-	79
	C	22	10	367	2	-	7	1	-	-	-	1	-	-	-	-	-	-	-	-	410
	D	2	1	6	27	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	14	1	6	1	-	22	4	2	8	-	-	-	-	-	-	-	-	-	-	56
	G	46	-	4	5	-	6	60	1	6	-	-	-	-	-	-	-	-	-	1	129
	H	12	1	1	-	-	1	6	3	4	-	-	-	-	-	-	-	-	-	1	29
	I	15	1	1	-	-	1	10	4	8	-	-	-	-	-	-	-	1	-	-	41
	J	1	-	-	-	-	-	1	-	-	0	-	-	-	-	-	-	-	-	1	3
	K	-	1	-	-	-	-	-	2	-	-	2	-	-	-	-	-	-	-	2	7
	L	1	-	-	-	-	-	-	-	1	-	-	0	-	-	-	-	-	-	-	2
	M	5	1	-	-	-	-	1	-	1	-	-	-	0	1	-	-	-	-	-	9
	N	8	-	1	-	-	-	3	3	2	-	1	-	-	0	-	-	-	-	-	18
	O	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	-	1	-	-	1	-	-	1	-	-	-	-	-	-	-	7	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	13	-	-	1	-	1	2	2	1	-	-	-	-	-	-	-	-	-	2	22
		289	72	419	41	0	47	115	24	38	0	4	0	1	1	0	0	8	0	9	1068

Table B.3: Discriminant matrix for independence Bayes (ψ_B, $P = P_I$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.364.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	116	8	16	4	1	5	38	5	8	1	-	-	1	-	1	-	-	-	5	209
	B	10	47	14	-	-	6	-	-	2	-	-	-	-	-	-	-	-	-	-	79
	C	21	9	364	3	2	6	1	1	-	-	1	-	-	-	1	-	-	-	1	410
	D	3	1	5	26	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	9	2	4	1	-	27	6	-	6	-	-	-	-	-	-	1	-	-	-	56
	G	35	-	4	6	1	4	67	2	7	-	-	-	1	-	-	-	-	-	2	129
	H	12	1	1	-	-	1	2	8	5	-	-	-	-	-	-	-	-	-	1	29
	I	14	1	-	-	-	1	10	4	10	-	-	-	-	-	-	-	1	-	-	41
	J	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	-	-	1	3
	K	-	1	-	-	-	-	-	2	-	-	3	-	-	-	-	-	-	-	2	7
	L	1	-	-	-	-	-	-	-	1	-	-	0	-	-	-	-	-	-	-	2
	M	4	1	-	-	-	-	1	-	2	-	-	-	0	1	-	-	-	-	-	9
	N	6	-	1	-	-	-	3	1	3	-	1	-	-	3	-	-	-	-	-	18
	O	-	-	-	-	-	-	-	-	-	-	-	-	1	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	1	-	1	-	-	-	-	-	-	0	-	-	-	2
	Q	1	1	-	-	-	-	-	-	1	-	-	-	-	-	-	-	8	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	8	-	-	1	-	1	4	3	1	2	-	-	-	-	-	-	-	-	3	22
		240	72	411	41	4	51	137	24	47	4	4	0	3	4	2	1	9	0	14	1068

Table B.4: Discriminant matrix for the small Bayesian network (ψ_B, $P = P_S$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.376.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	94	8	21	5	-	8	46	6	11	2	-	-	2	-	-	-	-	-	6	209
	B	7	45	15	-	1	7	1	-	3	-	-	-	-	-	-	-	-	-	-	79
	C	15	7	369	3	1	7	2	2	-	-	1	-	-	-	1	-	-	-	2	410
	D	2	1	5	26	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	9	-	5	1	-	30	5	-	6	-	-	-	-	-	-	-	-	-	-	56
	G	26	-	5	7	1	4	74	2	9	-	-	-	-	-	-	-	-	-	1	129
	H	10	1	1	-	-	1	3	6	4	-	-	-	-	-	-	-	1	-	2	29
	I	9	3	-	-	-	1	13	5	9	-	-	-	-	-	-	-	1	-	-	41
	J	1	-	-	-	-	-	1	-	-	0	-	-	-	-	-	-	-	-	1	3
	K	-	1	-	-	-	-	1	2	-	-	1	-	-	-	-	-	-	-	2	7
	L	-	-	-	-	-	-	-	1	1	-	-	0	-	-	-	-	-	-	-	2
	M	4	1	-	-	-	-	2	-	1	-	-	-	0	1	-	-	-	-	-	9
	N	4	-	1	-	-	-	4	2	3	-	1	-	-	3	-	-	-	-	-	18
	O	-	-	-	-	-	-	-	-	-	-	-	-	1	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	1	1	-	-	-	1	-	-	-	-	-	-	-	-	-	7	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	6	-	-	1	-	1	4	4	2	1	1	-	-	-	-	-	-	-	2	22
		188	68	425	43	3	59	163	30	49	3	4	0	3	4	1	0	9	0	16	1068

Table B.5: Discriminant matrix for the neural network (ψ_N, $n = 1$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.378.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	121	10	22	3	-	3	35	1	7	1	1	-	-	2	-	-	1	-	2	209
	B	5	40	25	1	-	6	-	-	2	-	-	-	-	-	-	-	-	-	-	79
	C	20	7	372	4	-	4	3	-	-	-	-	-	-	-	-	-	-	-	-	410
	D	1	-	9	22	-	1	5	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	2	1	13	1	-	30	5	1	3	-	-	-	-	-	-	-	-	-	-	56
	G	42	-	11	5	-	5	54	2	10	-	-	-	-	-	-	-	-	-	-	129
	H	16	1	1	-	-	2	5	2	2	-	-	-	-	-	-	-	-	-	-	29
	I	14	2	1	-	-	2	8	3	10	-	-	-	-	-	-	-	1	-	-	41
	J	1	-	1	-	-	-	1	-	-	0	-	-	-	-	-	-	-	-	-	3
	K	1	1	1	-	-	-	2	-	-	-	2	-	-	-	-	-	-	-	-	7
	L	1	-	-	-	-	-	-	-	1	-	-	0	-	-	-	-	-	-	-	2
	M	4	-	-	-	-	-	4	-	1	-	-	-	0	-	-	-	-	-	-	9
	N	9	-	1	-	-	-	3	1	2	-	-	-	-	2	-	-	-	-	-	18
	O	-	-	-	-	-	-	1	-	-	-	-	-	-	-	0	-	-	-	-	1
	P	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	2	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	7	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	9	-	3	1	-	-	4	-	2	-	-	-	-	1	-	-	-	-	2	22
		250	62	463	37	0	54	130	10	40	1	3	0	0	5	0	0	9	0	4	1068

Table B.6: Discriminant matrix for the inferential rule-based system (ψ_R, $Q = Q_I$, $T = T_I$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.382.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	128	4	38	-	-	-	24	6	3	-	-	-	-	-	-	-	3	3	-	209
	B	8	47	20	-	-	1	1	-	2	-	-	-	-	-	-	-	-	-	-	79
	C	20	21	356	1	-	9	1	2	-	-	-	-	-	-	-	-	-	-	-	410
	D	8	-	8	16	-	-	6	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	2	4	21	-	-	24	2	1	2	-	-	-	-	-	-	-	-	-	-	56
	G	39	1	7	6	-	1	61	5	4	-	-	2	-	1	-	2	-	-	-	129
	H	7	1	4	-	-	1	4	7	1	-	1	-	-	-	-	1	1	1	-	29
	I	9	-	4	-	-	4	7	4	7	-	-	-	-	2	-	-	2	2	-	41
	J	3	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	3
	K	2	-	-	-	-	-	1	-	-	-	4	-	-	-	-	-	-	-	-	7
	L	-	-	-	-	-	1	1	-	-	-	-	0	-	-	-	-	-	-	-	2
	M	4	-	-	-	-	1	1	-	-	-	-	-	0	1	1	-	-	1	-	9
	N	8	-	1	-	-	-	1	2	-	-	1	-	-	5	-	-	-	-	-	18
	O	-	-	-	-	-	-	-	-	-	-	-	-	1	-	0	-	-	-	-	1
	P	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	5	1	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	13	-	-	1	-	-	6	1	1	-	-	-	-	-	-	-	-	-	0	22
		253	80	463	24	0	42	117	28	20	0	5	2	1	9	1	3	11	8	0	1068

Table B.7: Discriminant matrix for the large Bayesian network (ψ_B, $P = P_L$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.383.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	100	7	24	2	-	6	43	6	13	-	-	-	3	-	-	-	-	-	5	209
	B	10	41	17	-	-	7	1	-	3	-	-	-	-	-	-	-	-	-	-	79
	C	10	5	366	6	1	10	3	4	1	-	1	-	-	-	1	-	-	-	2	410
	D	2	-	7	25	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	9	1	5	1	-	26	7	-	7	-	-	-	-	-	-	-	-	-	-	56
	G	32	-	6	6	-	4	71	2	8	-	-	-	-	-	-	-	-	-	-	129
	H	11	1	2	-	-	1	2	5	6	-	-	-	-	-	-	-	-	-	1	29
	I	11	1	-	-	-	1	10	6	11	-	-	-	-	-	-	-	1	-	-	41
	J	1	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	1	3
	K	1	1	-	-	-	-	1	1	-	-	2	-	-	-	-	-	-	-	1	7
	L	-	-	-	-	-	-	1	-	1	-	-	0	-	-	-	-	-	-	-	2
	M	3	1	-	-	-	-	3	-	2	-	-	-	0	-	-	-	-	-	-	9
	N	5	-	1	-	-	-	4	2	3	-	1	-	-	2	-	-	-	-	-	18
	O	-	-	-	-	-	-	-	-	-	-	-	-	1	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	-	1	-	-	1	-	-	1	-	-	-	-	-	-	-	7	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	5	-	1	1	-	1	5	4	1	1	1	-	-	-	-	-	-	-	2	22
		201	58	432	41	1	57	157	30	57	2	5	0	4	2	1	0	8	0	12	1068

Table B.8: Discriminant matrix for the flowchart program (ψ_F) when tested on all 1068 test cases. Error rate = 0.404.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	135	6	15	1	-	5	26	7	6	-	2	-	4	-	-	-	-	2	-	209
	B	6	42	24	-	-	3	-	-	4	-	-	-	-	-	-	-	-	-	-	79
	C	22	24	340	10	1	8	1	3	1	-	-	-	-	-	-	-	-	-	-	410
	D	12	-	8	15	-	1	2	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	4	7	8	1	1	24	1	1	8	-	-	-	1	-	-	-	-	-	-	56
	G	44	1	6	3	-	6	52	6	9	-	-	-	-	1	-	1	-	-	-	129
	H	3	1	-	-	-	3	4	14	4	-	-	-	-	-	-	-	-	-	-	29
	I	15	1	-	-	-	3	7	6	8	-	-	-	-	-	-	-	-	1	-	41
	J	2	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	3
	K	2	-	-	-	-	-	1	-	-	-	0	-	1	-	-	-	-	3	-	7
	L	-	-	-	-	-	1	1	-	-	-	-	0	-	-	-	-	-	-	-	2
	M	4	-	-	-	-	1	2	-	-	-	-	-	1	-	1	-	-	-	-	9
	N	6	-	1	-	-	-	1	3	2	-	-	-	1	4	-	-	-	-	-	18
	O	-	-	-	-	-	-	-	-	-	-	-	-	1	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	1	-	1	-	-	-	-	-	-	0	-	-	-	2
	Q	7	2	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9	-	0
	S	15	-	-	-	-	-	4	1	2	-	-	-	-	-	-	-	-	-	0	22
		277	84	404	30	2	55	103	41	46	1	2	0	9	5	1	1	1	6	0	1068

Table B.9: Discriminant matrix for iterative partitioning (ψ_I, $\alpha = 26.0$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.417.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	147	4	12	1	-	2	34	4	2	-	-	-	-	-	-	-	3	-	-	209
	B	13	44	17	-	-	4	-	-	-	-	-	-	-	-	-	-	1	-	-	79
	C	28	11	351	4	-	7	4	-	1	-	-	-	-	-	-	-	4	-	-	410
	D	6	-	21	5	-	-	6	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	1	-	0	-	-	-	-	-	-	-	-	-	-	-	1	-	-	2
	F	18	1	11	-	-	19	4	-	1	-	-	-	-	-	-	-	2	-	-	56
	G	61	-	6	2	-	3	50	4	2	-	-	-	-	-	-	-	1	-	-	129
	H	15	-	1	-	-	1	9	1	-	-	2	-	-	-	-	-	-	-	-	29
	I	28	1	-	-	-	1	6	5	0	-	-	-	-	-	-	-	-	-	-	41
	J	1	-	1	-	-	-	1	-	-	0	-	-	-	-	-	-	-	-	-	3
	K	2	-	-	-	-	-	-	-	-	-	5	-	-	-	-	-	-	-	-	7
	L	2	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	2
	M	3	-	-	-	-	2	3	-	-	-	-	-	0	-	-	-	-	-	1	9
	N	7	-	1	-	-	-	6	1	-	-	3	-	-	0	-	-	-	-	-	18
	O	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	1
	P	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	1	-	-	2
	Q	9	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	9	-	2	2	-	-	8	-	-	-	1	-	-	-	-	-	-	-	0	22
		351	62	424	14	0	39	131	15	6	0	11	0	0	0	0	0	14	0	1	1068

Table B.10: Discriminant matrix for the exemplar model (ψ_E) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.428.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	71	7	19	4	1	7	31	3	35	2	1	1	3	7	-	1	9	-	7	209
	B	3	49	16	1	-	4	2	-	1	-	-	-	-	1	-	-	2	-	-	79
	C	12	11	369	4	3	4	3	-	3	-	-	-	-	-	-	-	-	-	1	410
	D	-	-	9	25	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	3	1	6	4	1	17	9	-	6	-	-	1	-	2	-	2	1	-	1	56
	G	19	1	6	9	-	4	54	3	18	-	-	1	4	4	-	1	-	-	5	129
	H	5	-	1	-	-	3	7	3	7	-	1	-	-	2	-	-	-	-	-	29
	I	5	1	-	-	-	2	7	2	13	-	-	-	-	7	-	-	3	-	1	41
	J	1	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	2	3
	K	1	1	-	-	-	-	1	1	-	-	1	-	-	-	-	-	-	-	2	7
	L	-	-	-	-	-	-	-	-	1	-	-	0	-	1	-	-	-	-	-	2
	M	1	-	1	-	-	1	2	-	2	-	-	-	0	1	-	-	-	1	-	9
	N	6	-	2	-	-	-	3	1	3	-	-	-	1	2	-	-	-	-	-	18
	O	-	-	-	-	-	-	-	-	-	-	-	-	1	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	2	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	7	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	9	-	3	2	-	-	5	-	1	-	1	-	-	1	-	-	-	-	0	22
		138	72	437	49	5	42	130	13	90	2	4	3	9	28	0	4	22	1	19	1068

Table B.11: Discriminant matrix for nearest neighbours using the Hamming metric (ψ_K, $\delta = \delta_f$, $k = 21$) when tested on all 1068 test cases using a 'leave-out-101' training strategy. Error rate = 0.485.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	122	3	70	-	-	1	11	-	2	-	-	-	-	-	-	-	-	-	-	209
	B	8	8	63	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	79
	C	12	-	398	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	410
	D	1	-	33	3	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	17	1	32	-	-	0	3	-	3	-	-	-	-	-	-	-	-	-	-	56
	G	75	-	26	1	-	1	19	-	7	-	-	-	-	-	-	-	-	-	-	129
	H	17	1	7	-	-	-	2	0	2	-	-	-	-	-	-	-	-	-	-	29
	I	34	1	3	-	-	1	2	-	0	-	-	-	-	-	-	-	-	-	-	41
	J	-	-	2	-	-	-	1	-	-	0	-	-	-	-	-	-	-	-	-	3
	K	4	1	1	-	-	-	1	-	-	-	0	-	-	-	-	-	-	-	-	7
	L	2	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	2
	M	7	-	1	-	-	-	1	-	-	-	-	-	0	-	-	-	-	-	-	9
	N	13	-	4	-	-	-	1	-	-	-	-	-	-	0	-	-	-	-	-	18
	O	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	1
	P	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	3	7	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	11	-	8	1	-	-	2	-	-	-	-	-	-	-	-	-	-	-	0	22
		326	18	657	5	0	3	45	0	14	0	0	0	0	0	0	0	0	0	0	1068

Table B.12: Discriminant matrix for independence Bayes (ψ_B, $P = P_I$) when tested on all 1068 test cases and trained on 10^6 cases generated from the causal rule-based system (Q_C, T_C), using a 'leave-out-101' training strategy. Error rate = 0.364.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	97	6	22	4	-	4	48	7	7	1	2	-	2	2	-	-	3	-	4	209
	B	7	46	13	-	-	6	2	-	5	-	-	-	-	-	-	-	-	-	-	79
	C	11	11	368	1	4	8	3	-	1	-	-	-	-	-	-	-	1	-	2	410
	D	3	-	7	27	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	38
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	6	2	6	1	-	27	11	1	2	-	-	-	-	-	-	-	-	-	-	56
	G	21	-	5	9	-	2	78	5	6	-	-	2	-	-	-	-	-	-	1	129
	H	11	-	1	-	1	2	6	9	5	-	-	-	-	-	-	-	-	-	-	29
	I	10	-	-	-	-	1	13	1	15	-	-	-	-	-	-	-	1	-	-	41
	J	2	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	3
	K	1	1	-	-	-	-	1	-	-	-	4	-	-	-	-	-	-	-	-	7
	L	-	-	-	-	-	-	-	-	2	-	-	0	-	-	-	-	-	-	-	2
	M	3	-	1	-	-	-	4	-	1	-	-	-	0	-	-	-	-	-	-	9
	N	3	-	1	-	-	-	7	-	1	-	3	-	-	2	-	-	-	-	1	18
	O	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	10	-	-	11
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	5	-	-	3	-	2	7	2	1	-	1	-	-	-	-	-	-	-	1	22
		181	66	426	45	5	52	184	19	46	2	10	2	2	4	0	0	15	0	9	1068

Table B.13: Discriminant matrix for the neural network (ψ_N, $n = 1$) when tested on all 751 test cases with definite diagnoses. See text for description of cross-validation strategy. Error rate = 0.269.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	135	7	12	-	-	5	6	1	8	-	1	-	-	-	-	-	-	-	2	177
	B	3	38	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54
	C	5	3	309	3	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	322
	D	2	-	8	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	5	-	9	1	-	33	2	-	3	-	-	-	-	-	-	-	-	-	-	53
	G	18	-	5	-	-	6	13	-	1	-	-	-	-	-	-	-	-	-	-	43
	H	9	1	1	-	-	2	2	1	1	-	-	-	-	-	-	-	-	-	-	17
	I	12	-	-	-	-	1	3	1	2	-	-	-	-	1	-	-	-	-	-	20
	J	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	0
	K	1	-	-	-	-	-	2	-	-	-	1	-	-	-	-	-	-	-	-	4
	L	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	0
	M	3	-	-	-	-	-	1	-	-	-	-	-	0	-	-	-	-	-	-	4
	N	6	-	1	-	-	-	2	-	1	-	-	-	-	1	-	-	-	-	-	11
	O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	0
	P	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	0	-	-	2
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	10	-	1	1	-	1	-	3	-	-	-	-	-	-	-	-	-	-	0	16
		211	49	359	21	0	49	33	3	20	0	2	0	0	2	0	0	0	0	2	751

Table B.14: Discriminant matrix for nearest neighbours using the Bayes metric ($\psi_K, \delta = \delta_b$, $k = 48$) when tested on all 751 test cases with definite diagnoses. See text for description of cross-validation strategy. Error rate = 0.273.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	150	3	11	1	-	3	9	-	-	-	-	-	-	-	-	-	-	-	-	177
	B	4	39	10	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	54
	C	5	8	297	4	-	8	-	-	-	-	-	-	-	-	-	-	-	-	-	322
	D	1	-	2	19	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	24
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	19	-	3	1	-	26	2	-	2	-	-	-	-	-	-	-	-	-	-	53
	G	21	-	2	-	-	5	15	-	-	-	-	-	-	-	-	-	-	-	-	43
	H	13	-	1	-	-	1	1	0	1	-	-	-	-	-	-	-	-	-	-	17
	I	13	-	1	-	-	2	4	-	0	-	-	-	-	-	-	-	-	-	-	20
	J	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	0
	K	1	-	3	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	4
	L	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	0
	M	1	-	1	-	-	-	2	-	-	-	-	-	0	-	-	-	-	-	-	4
	N	8	-	1	-	-	-	1	-	1	-	-	-	-	0	-	-	-	-	-	11
	O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	0
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	2
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	14	-	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	0	16
		250	50	336	26	0	47	38	0	4	0	0	0	0	0	0	0	0	0	0	751

Table B.15: Discriminant matrix for the inferential rule-based system ($\psi_R, Q = Q_I, T = T_I$) when tested on all 751 test cases with definite diagnoses. See text for description of cross-validation strategy. Error rate = 0.276.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	129	1	28	-	-	4	5	6	1	-	-	-	-	-	-	-	-	3	-	177
	B	3	42	5	-	-	2	-	-	2	-	-	-	-	-	-	-	-	-	-	54
	C	6	2	305	2	-	7	-	-	-	-	-	-	-	-	-	-	-	-	-	322
	D	4	-	4	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24
	E	-	-	2	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	F	3	-	19	2	-	27	1	-	1	-	-	-	-	-	-	-	-	-	-	53
	G	18	-	2	1	-	2	14	1	3	-	-	-	-	1	-	1	-	-	-	43
	H	4	-	3	-	-	1	3	3	1	-	1	-	-	-	-	-	1	-	-	17
	I	6	-	2	-	-	2	2	3	3	-	-	-	-	1	-	-	-	1	-	20
	J	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	0
	K	1	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	1	-	4
	L	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	0
	M	3	-	-	-	-	-	1	-	-	-	-	-	0	-	-	-	-	-	-	4
	N	4	-	1	-	-	-	1	2	-	-	1	-	-	2	-	-	-	-	-	11
	O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	0
	P	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-	2
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	0
	S	10	-	-	1	-	-	2	2	1	-	-	-	-	-	-	-	-	-	0	16
		193	45	371	22	0	45	29	17	12	0	4	0	0	4	0	1	2	6	0	751

Table B.16: Discriminant matrix for independence Bayes (ψ_B, $P = P_I$) when tested on all 751 test cases with definite diagnoses. See text for description of cross-validation strategy. Error rate = 0.289.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Actual Diagnosis	A	**130**	4	9	1	-	5	17	2	1	1	-	1	1	-	-	-	-	-	5
	B	3	**40**	8	-	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-
	C	6	8	**293**	3	1	6	1	-	-	-	1	1	-	-	-	-	-	-	-
	D	2	-	2	**19**	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
	E	-	-	2	-	**0**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	F	13	-	3	1	-	**24**	4	-	6	-	-	-	-	-	-	2	-	-	-
	G	16	-	2	-	-	3	**19**	-	3	-	-	-	-	-	-	-	-	-	-
	H	7	-	1	-	-	2	2	**2**	2	-	-	-	-	-	-	-	-	-	1
	I	9	-	-	-	-	2	6	-	1	-	-	-	-	-	-	-	1	-	1
	J	-	-	-	-	-	-	-	-	-	**0**	-	-	-	-	-	-	-	-	-
	K	-	-	-	-	-	-	-	-	-	-	**4**	-	-	-	-	-	-	-	-
	L	-	-	-	-	-	-	-	-	-	-	-	**0**	-	-	-	-	-	-	-
	M	1	-	1	-	-	-	2	-	-	-	-	-	**0**	-	-	-	-	-	-
	N	4	-	1	-	-	-	2	-	3	-	-	1	-	**0**	-	-	-	-	-
	O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**0**	-	-	-	-
	P	-	-	-	-	-	-	1	-	1	-	-	-	-	-	-	**0**	-	-	-
	Q	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1	-	-
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**0**	-
	S	8	-	1	1	-	-	4	1	-	-	-	-	-	-	-	-	-	-	**1**
		201	52	323	25	2	44	59	5	18	1	5	3	1	0	0	2	2	0	8

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Actual Diagnosis	A	**114**	1	21	1	-	5	21	3	5	-	1	-	-	-	-	-	2	-	3
	B	5	**40**	5	-	-	1	-	-	3	-	-	-	-	-	-	-	-	-	-
	C	2	6	**306**	-	2	5	-	-	-	-	-	-	-	-	-	-	-	-	1
	D	1	-	5	**17**	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
	E	-	-	2	-	**0**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	F	10	-	6	1	-	**27**	9	-	-	-	-	-	-	-	-	-	-	-	-
	G	9	-	2	1	-	2	**21**	4	2	-	-	-	-	-	-	1	-	-	1
	H	8	-	1	-	1	1	3	**0**	3	-	-	-	-	-	-	-	-	-	-
	I	6	-	-	-	-	2	8	-	**3**	-	-	-	-	-	-	-	1	-	-
	J	-	-	-	-	-	-	-	-	-	**0**	-	-	-	-	-	-	-	-	-
	K	1	-	-	-	-	-	1	-	-	-	**2**	-	-	-	-	-	-	-	-
	L	-	-	-	-	-	-	-	-	-	-	-	**0**	-	-	-	-	-	-	-
	M	2	-	1	-	-	-	-	-	1	-	-	-	**0**	-	-	-	-	-	-
	N	5	-	1	-	-	-	4	1	-	-	-	-	-	**0**	-	-	-	-	-
	O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**0**	-	-	-	-
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	**0**	-	-	-
	Q	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1	-	-
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**0**	-
	S	7	-	1	2	-	1	4	1	-	-	-	-	-	-	-	-	-	-	**0**
		170	47	351	22	3	44	75	9	17	0	3	0	0	0	0	1	4	0	5

Table B.17: Discriminant matrix for nearest neighbours with the Bayes metric (ψ_K, $\delta = \delta_b$, $k = 300$) when tested on all 751 test cases with definite diagnoses, and trained on 10^6 cases generated from the causal rule-based system (Q_C, T_C). See text for description of cross-validation strategy. Error rate = 0.293.

Table B.18: Discriminant matrix for nearest neighbours using the Bayes metric ($\psi_K, \delta = \delta_b$, $k = 29$) when tested on all 1270 cases in the database using a 'leave-out-one' training strategy. Error rate = 0.370.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	156	9	23	8	-	4	32	4	11	-	-	-	-	-	-	-	1	-	5	253
	B	15	58	14	1	-	3	3	1	-	-	-	-	-	-	-	-	-	-	-	95
	C	22	10	424	2	-	9	1	-	-	-	-	-	-	-	-	-	-	-	-	468
	D	3	1	6	35	-	1	2	-	-	-	-	-	-	-	-	-	-	-	-	48
	E	-	-	3	-	0	1	-	-	-	-	-	-	-	-	-	-	-	-	-	4
	F	19	1	9	-	-	27	7	2	6	-	-	-	-	-	-	-	-	-	1	72
	G	44	-	4	8	-	3	79	3	6	-	-	-	-	-	-	-	-	-	2	149
	H	16	3	-	-	-	2	9	2	6	-	-	-	-	-	-	-	-	-	1	39
	I	19	1	1	-	-	3	14	7	8	-	-	-	-	-	-	-	1	-	-	54
	J	3	-	-	-	-	-	1	-	-	0	-	-	-	-	-	-	-	-	1	5
	K	3	1	-	-	-	-	-	2	-	-	0	-	-	-	-	-	-	-	2	8
	L	-	-	-	-	-	-	1	-	2	-	-	0	-	-	-	-	-	-	-	3
	M	8	1	-	-	-	-	1	-	2	-	-	-	0	-	-	-	-	-	-	12
	N	10	-	1	-	-	-	4	4	1	-	-	-	-	0	-	-	-	-	-	20
	O	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	0	-	-	-	2
	Q	1	1	1	-	-	-	-	-	1	-	-	-	-	-	-	-	9	-	1	14
	R	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	1
	S	9	-	1	1	-	1	4	3	1	-	-	-	-	-	-	-	-	-	2	22
		330	86	487	55	0	54	160	28	44	0	0	0	0	0	0	0	11	0	15	1270

Table B.19: Discriminant matrix for independence Bayes (ψ_B, $P = P_I$) when tested on all 1270 cases in the database using a 'leave-out-one' training strategy. Error rate = 0.374.

		Computer Diagnosis																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
Actual Diagnosis	A	133	9	21	8	-	5	42	7	18	1	-	-	1	2	1	-	1	-	4	253
	B	12	59	16	-	-	6	-	-	2	-	-	-	-	-	-	-	-	-	-	95
	C	21	10	421	3	2	7	2	-	-	-	-	-	-	1	1	-	-	-	-	468
	D	4	1	4	34	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	48
	E	-	-	3	-	0	1	-	-	-	-	-	-	-	-	-	-	-	-	-	4
	F	19	2	5	1	-	29	7	-	7	-	-	-	-	-	-	1	-	-	1	72
	G	33	-	3	10	1	5	81	3	8	-	-	1	2	-	-	-	-	-	2	149
	H	14	2	1	-	-	3	4	7	7	-	-	-	-	-	-	-	-	-	1	39
	I	17	1	-	-	-	4	10	9	12	-	-	-	-	-	-	-	1	-	-	54
	J	-	-	-	-	-	-	1	-	-	3	-	-	-	-	-	-	-	-	1	5
	K	-	1	-	-	-	-	-	2	-	-	2	-	-	1	-	-	-	-	2	8
	L	-	-	-	-	-	-	2	-	-	-	1	0	-	-	-	-	-	-	-	3
	M	5	1	-	-	-	-	2	-	2	-	-	-	0	1	1	-	-	-	-	12
	N	6	-	1	-	-	-	4	2	3	-	1	-	-	2	-	-	-	-	1	20
	O	-	-	-	-	-	-	-	-	-	-	-	-	1	-	0	-	-	-	-	1
	P	-	-	-	-	-	-	1	-	1	-	-	-	-	-	-	0	-	-	-	2
	Q	1	1	-	-	-	-	-	-	1	-	-	-	-	-	-	-	10	-	1	14
	R	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	1
	S	6	-	-	1	-	1	5	4	1	2	-	-	-	-	-	-	-	-	2	22
		272	87	475	57	3	61	166	34	62	6	4	1	4	7	3	1	12	0	15	1270

Bibliography

[Aas93] **Aase O, Jonsbu J, Liestøl K, Rollag A, Eriksen J.** Decision support by computer analysis of selected case history variables in the emergency room among patients with acute chest pain. *European Heart Journal* (1993) **14** 433-40.

[Ada86] **Adams ID, Chan M, Clifford PC, Cooke WM, Dallos V, de Dombal FT, Edwards MH, Hancock DM, Hewett DJ, McIntyre N, Somerville PG, Spiegelhalter DJ, Wellwood J, Wilson DH.** Computer-aided diagnosis of acute abdominal pain: a multicentre study. *British Medical Journal* (1986) **293** 800-4.

[Aka92] **Akay M.** Noninvasive diagnosis of coronary artery disease using a neural network algorithm. *Biological Cybernetics* (1992) **67** 361-7.

[And82] **Anderson JA.** Logistic discrimination. In **Krishnaiah PR, Kanal LN (eds).** *Handbook of Statistics 2.* North-Holland: Amsterdam, New York, Oxford (1982) 169-91.

[And91] **Andreassen S, Jensen FV, Olesen KG.** Medical expert systems based on causal probabilistic networks. *International Journal of Biomedical Computing* (1991) **28** *i* 1-30.

[Bax91] **Baxt WG.** Use of an artificial neural network for the diagnosis of myocardial infarction. *Annals of Internal Medicine* (1991) **115** *xi* 843-8.

[Bez91] **Bezem J.** Semantics and consistency of rule-based expert systems. *Journal of Logic and Computation* (1991) **1** *iv* 477-95.

[Bou90] **Bounds DG, Lloyd PJ, Mathew BG.** A comparison of neural network and other pattern recognition approaches to the diagnosis of low back pain. *Neural Networks* (1990) **3** 583-91.

[Bre84] **Breiman L, Friedman JH, Olshen RA, Stone CJ.** *Classification and Regression Trees.* Wadsworth and Brooks/Cole: Monterey (1984).

[Ces90] **Cestnik B.** Estimating probabilities: a crucial task in machine learning. *Proceedings of the 9th European Conference on Artificial Intelligence (ECAI 90).* Pitman: London (1990) 147-9.

[Cha84] **Chamberlain G.** *Contemporary Gynaecology.* Butterworths: London, Boston etc. (1984).

[Cha87] **Chard T.** Self-learning for a Bayesian knowledge base: how long does it take for the machine to educate itself? *Methods of Information in Medicine* (1987) **26** *iv* 185-8.

[Cha88] **Chandy KM, Misra J.** *Parallel Program Design — A Foundation.* Addison-Wesley: Reading, Menlo Park etc. (1988).

[Cha89] **Chard T, Rubenstein EM.** A model-based system to determine the relative value of different variables in a diagnostic system using Bayes theorem. *International Journal of Biomedical Computing* (1989) **24** *ii* 133-42.

[Cla92] **Clark LA, Pregibon D.** Tree-based models. In **Chambers JM, Hastie TJ (eds).** *Statistical models in S.* Wadsworth and Brooks/Cole: Pacific Grove (1992).

[Coo86] **Cooper GF.** A diagnostic method that uses causal knowledge and linear programming in the application of Bayes' formula. *Computer Methods and Programs in Biomedicine* (1986) **22** 223-37.

[Coo89] **Cooper GF.** Current research directions in the development of expert systems based on belief networks. *Applied Stochastic Models and Data Analysis* (1989) **5** *i* 39-52.

[Cor86] **Corlett RA and Todd SJ.** A Monte Carlo approach to uncertain inference. In **Cohn AG, Thomas JR (eds.).** *Artificial Intelligence and its Applications.* John Wiley & Sons: Chichester (1986) 127-37.

[Cri87] **Crichton NJ, Fryer JG, Spicer CC.** Some points on the use of 'independent Bayes' to diagnose acute abdominal pain. *Statistics in Medicine* (1987) **6** *viii* 945-59.

[Cri89] **Crichton NJ, Hinde JP.** Correspondence analysis as a screening method for indicants for clinical diagnosis. *Statistics in Medicine* (1989) **8** *xi* 1351-62.

[Cro74] **Croft DJ, Machol RE.** Mathematical methods in medical diagnosis. *Annals of Biomedical Engineering* (1987) **2** 69-89.

[Cun89] **Cunningham FG, MacDonald PC, Gant NF.** *Williams Obstetrics.* 18th ed. Prentice-Hall International Inc: London (1989).

[Dan92] **Dan Q, Dudeck J.** Certainty factor theory: its probabilistic interpretations and problems. *Artificial Intelligence in Medicine* (1992) **4** *i* 21-34.

[Dav71] **Davis R, Buchanan B, Shortliffe E.** Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence* (1977) **8** *i* 15-45.

[Dic88] **Dickson JAS, Jones A, Telfer S, de Dombal FT.** Acute abdominal pain in children. *Scandinavian Journal of Gastroenterology (Supplements)* (1988) **144** 43-6.

[Dix91] **Dixon JM, Elton RA, Rainey JB, Macleod DAD.** Rectal examination in patients with pain in the right lower quadrant of the abdomen. *British Medical Journal* (1991) **302** 386-8.

[Dom72] **De Dombal FT, Leaper DJ, Horrocks JC, Staniland JR, McCann AP.** Computer-aided diagnosis of acute abdominal pain. *British Medical Journal* (1972) *ii* 9-13.

[Dom78] **De Dombal FT.** Medical diagnosis from a clinician's point of view. *Methods of Information in Medicine* (1978) **17** *i* 28-35.

[Dom84] **De Dombal FT.** Clinical decision making and the computer: consultant, expert, or just another test? *British Journal of Healthcare Computing* (1984) **1** *i* 7-14.

[Dom89] **De Dombal FT.** Computer-aided decision support in clinical medicine. *International Journal of Biomedical Computing* (1989) **24** *i* 9-16.

[Dom90] **De Dombal FT.** Computer-aided decision support in acute abdominal pain, with special refereuce to the EC concerted action. *International Journal of Biomedical Computing* (1990) **26** *iii* 183-8.

[Dom91] **De Dombal FT.** The diagnosis of acute abdominal pain with computer assistance: worldwide perspective. *Annales de Chirurgie* (1991) **45** *iv* 273-7.

[Dom92] **De Dombal FT, Barnes S, Dallos V, Kumar PS, Sloan J, Chan M, Stapleton C, Wardle KS.** How should computer-aided decision-support systems present their predictions to the practising surgeon? *Theoretical Surgery* (1992) **7** 111-6.

[Dud79] **Duda R, Gaschnig J, Hart P.** Model design in the PROSPECTOR consultant system for mineral exploration. In **Michie D (ed).** *Expert Systems in the Micro-Electronic Age.* Edinburgh University Press: Edinburgh (1979) 153-67.

[Ebe91] **Ebehart RC, Dobbins RW, Hutton LV.** Neural network paradigm comparisons for appendicitis diagnoses. *Proceedings of 4th Annual IEEE Symposium on Computer-Based Medical Systems* (1991) 298-304.

[Edw84] **Edwards FH, Davies RS.** Use of a Bayesian algorithm in the computer-assisted diagnosis of appendicitis. *Surgery, Gynecology and Obstetrics* (1984) **158** 219-22.

[Emp88] **Emparanza JI, Aldamiz-Echevarria L, Perez-Yarza EG, Aldazabal P, Tovar J.** Statistical approach to the acute abdomen diagnosis in children. *Expert Systems and Decision Support in Medicine. 33rd Annual Meeting of the GMDS EFMI Special Topic Meeting. Peter L. Reichertz Memorial Conference.* (1988) 131-5.

[Eng92] **Engle RL.** Attempts to use computers as diagnostic aids in medical decision making: a thirty-year experience. *Perspectives in Biology and Medicine* (1992) **35** *ii* 207-19.

[Fen90] **Fenyô G.** Computer-aided diagnosis and decision-making in acute abdominal pain. *Digestive Diseases* (1990) **8** *iii* 639-65.

[Fen87] **Fenyö G, Clamp SE, de Dombal FT, Engström L, Hedlund M, Leijonmarck C, Wilczek H.** Computer-aided diagnosis of 233 acute abdominal cases at Nacka Hospital Sweden. *Scandinavian Journal of Gastroenterology (Supplements)* (1987) **128** 178-8.

[Fla92] **Flamant Y, The ARC & AURC Cooperative Group.** L'aide au diagnostic par ordinateur. Un examen complémentaire très "cliniqne". *Revue du Praticien (Paris)* (1992) **42** *vi* 694-6.

[Fox80] **Fox J, Barber D, Bardhan KD.** A quantitative comparison with rule-based diagnostic inference. *Methods of Information in Medicine* (1980) **19** *iv* 210-5.

[Fra91] **Franklin RCG, Spiegelhalter DJ, Macartney FJ, Bull K.** Evaluation of a diagnostic algorithm for heart disease in neonates. *British Medical Journal* (1991) **302** 935-9.

[Fri86] **Fries JF, McShane DJ.** ARAMIS (The American Rheumatism Association Medical Information System) — a prototypical national chronic-disease data bank. *The Western Journal of Medicine* (1986) **145** 798-804.

[Fry78] **Fryback DG.** Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research* (1978) **11** *v* 423-34.

[Gam91] **Gammerman A, Thatcher AR.** Bayesian diagnostic probabilities without assuming independence of symptoms. *Methods of Information in Medicine* (1991) **30** *i* 15-22.

[Gil73] **Gill PW, Leaper DJ, Guillou PJ, Staniland JR, Horrocks JC, de Dombal FT.** Observer variation in clinical diagnosis — a computer-aided assessment of its magnitnde and importance in 552 patients with abdominal pain. *Methods of Information in Medicine* (1973) **12** *ii* 108-13.

[Gil91] **Gillmer MDG, Steer PJ, Woolfson J.** *100 Case Histories in Obstetrics and Gynaecology.* Churchill Livingstone: Edinburgh, London etc. (1991).

[Goo85] **Goodall A.** *The Guide to Expert Systems.* Learned Information: Oxford, New Jersey (1985).

[Gro90] **Gross GW, Boone JM, Greco-Hunt V, Greenberg B.** Neural networks in radiologic diagnosis II. Interpretation of neonatal chest radiographs. *Investigative Radiology* (1990) **25** *ix* 1017-23.

[Gun91] **Gunn AA.** The acute abdomen: the role of computer-assisted diagnosis. *Baillière's Clinical Gastroenterology* (1991) **5** *iii* 639-65.

[Gup89] **Guppy KH, Detrano R, Ahbassi N, Janosi A, Sandhu S, Froelicher V.** The reliability of probability analysis in the prediction of coronary artery disease in two hospitals. *Medical Decision Making* (1989) **9** *iii* 181-9.

[Hai88] **Hains G, Todd BS.** The parallel implementation of a medical diagnostic model. *Proceeding of the 3rd International Conference on Supercomputing* (1988) *i* 222-9.

[Haj85] **Hájek P.** Combining functions for certainty degrees in consulting systems. *International Journal of Man-Machine Studies* (1985) **22** *i* 59-76.

[Har89] **Hart A, Wyatt J.** Connectionist models in medicine: an investigation of their potential. *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine (AIME 89)* (1989) 115-24.

[Hec86] **Heckerman DE.** Probabilistic interpretations for MYCIN's certainty factors. In **Kanal LN, Lemmer JF (eds).** *Uncertainty in Artificial Intelligence.* North-Holland: Amsterdam, New York, Oxford (1986) 167-96.

[Hec88] **Heckerman DE, Horvitz EJ.** The myth of modularity in rule-based systems. In **Lemmer JF, Kanal LN (eds).** *Uncertainty in Artificial Intelligence 2.* North-Holland: Amsterdam, New York, Oxford (1988) 23-34.

[Hec92a] **Heckerman DE, Horvitz EJ, Nathwani BN.** Toward normative expert systems: Part I. The Pathfinder project. *Methods of Information in Medicine* (1992) **31** *ii* 90-105.

[Hec92b] **Heckerman DE, Nathwani BN.** Toward normative expert systems: Part II. Probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in Medicine* (1992) **32** *ii* 106-16.

[Hec92c] **Heckerman DE, Nathwani BN.** An evaluation of the diagnostic accuracy of Pathfinder. *Computers and Biomedical Research* (1992) **25** *i* 56-74.

[Hil84] **Hilden J.** Statistical diagnosis based on conditional independence does not require it. *Computers in Biology and Medicine* (1984) **14** *iv* 429-35.

[Hud91] **Hudson DL, Cohen ME, Anderson MF.** Computer-assisted differential diagnosis and management. *Proceedings of the 24th Annual Hawaii International Conference on System Sciences* (1991) **3** 218-26.

[Jac86] **Jackson P.** *Introduction to Expert Systems.* Addison-Wesley: Reading (1985).

[Kir87] **Kirkeby OJ, Risø C.** Use of a computer system for diagnosing acute abdominal pain in a small hospital. *Scandinavian Journal of Gastroenterology (Supplements)* (1987) **128** 174-6.

[Kni85] **Knill-Jones RP.** A formal approach to symptoms in dyspepsia. *Clinics in Gastroenterology* (1985) **14** *iii* 517-29.

[Lau88] **Lauritzen SL, Spiegelhalter DJ.** Local computation with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society (Series B)* (1988) **50** *ii* 157-224.

[Lav90] **Lavelle SM, Kanagaratnam B.** The informational value of clinical data. *International Journal of Biomedical Computing* (1990) **26** *iii* 203-9.

[Lea72] **Leaper DJ, Horrocks JC, Staniland JR, De Dombal FT.** Computer-assisted diagnosis of abdominal pain using "estimates" provided by clinicians. *British Medical Journal* (1972) *iv* 350-4.

[Led59] **Ledley RS, Lusted LB.** Reasoning foundations of medical diagnosis. *Science* (1959) **130** *i* 9-21.

[Lip61] **Lipkin M, Engle RL, Davis BJ, Zworykin VK, Ebald R, Sendrow M, Berkley C.** Digital computer as an aid to differential diagnosis. *Archives of Internal Medicine* (1961) **108** 56-72.

[Lip87] **Lippmann R.** An introduction to computing with neural nets. *IEEE Acoustics, Speech and Signal Processing Society Magazine* (1987) *April* 4-22.

[Low90] **Lowe D, Webb AR.** Exploiting prior knowledge in network optimization: an illustration from medical prognosis. *Network* (1990) **1** 299-323.

[Lud83] **Ludwig D, Heilbronn D.** The design and testing of a new approach to computer-aided differential diagnosis. *Methods of Information in Medicine* (1983) **22** *iii* 156-66.

[Mac78] **Macartney FJ.** Diagnostic logic. *British Medical Journal* (1987) **295** 1325-31.

[Mac91] **Maclin PS, Dempsey J, Brooks J, Rand J.** Using neural networks to diagnose cancer. *Journal of Medical Systems* (1991) **15** *i* 11-21.

[Mai88] **Maitra AK, Briggs PJ, McGeehan D.** Computer-unaided diagnosis of acute abdominal pain in an accident and emergency department. *Archives of Emergency Medicine* (1988) **5** 74-8.

[Mos52] **Mosteller F.** Some statistical problems in measuring the subjective response to drugs. *Biometrics* (1952) **8** *i* 220-6.

[Mul90] **Mulsant BH.** A neural network as an approach to clinical diagnosis. *MD Computing* (1990) **7** *i* 25-36.

[Nea89] **Neapolitan RE.** *Probabilistic Reasoning in Expert Systems.* John Wiley: New York, Chichester etc. (1989).

[Nor71] **Nordyke R, Kulikowski CA, Kulikowski CW.** A comparison of methods for the automated diagnosis of thyroid dysfunction. *Computers and Biomedical Research* (1971) **4** *iv* 374-89.

[Nor75a] **Norusis MJ, Jacques JA.** Diagnosis. I. Symptom nonindependence in mathematical models for diagnosis. *Computers and Biomedical Research* (1975) **8** *ii* 156-72.

[Nor75b] **Norusis MJ, Jacquez JA.** Diagnosis. II. Diagnostic models based on attribute clusters: a proposal and comparisons. *Computers and Biomedical Research* (1975) **8** *ii* 173-88.

[Ohm86] **Ohmann C, Künneke M, Zaczyk R, Thon K, Lorenz W.** Selection of variables using 'independence Bayes' in computer-aided diagnosis of upper gastrointestinal bleeding. *Statistics in Medicine* (1986) **5** *v* 503-15.

[Ohm88] **Ohmann C, Yang Q, Künneke M, Stöltzing H, Thon K, Lorenz W.** Bayes theorem and conditional dependence of symptoms: different models applied to data of upper gastrointestinal bleeding. *Methods of Information in Medicine* (1988) **27** *ii* 73-83.

[Ori86] **Orient JM.** Evaluation of abdominal pain: clinicians' performance compared with three protocols. *Southern Medical Journal* (1986) **79** *vii* 793-9.

[Pat89] **Paterson-Brown S, Vipond MN, Simms K, Gatzen C, Thompson JN, Dudley HAF.** Clinical decision making and laparoscopy versus computer prediction in the management of the acute abdomen. *British Journal of Surgery* (1989) **76** *x* 1011-3.

[Pat91] **Paterson-Brown S.** Strategies for reducing inappropriate laparotomy rate in the acute abdomen. *British Medical Journal* (1991) **303** 1115-8.

[Pau82] **Pauerstein CJ.** *Gynecological Disorders. Differential Diagnosis and Therapy.* Grune and Stratton: New York, London etc. (1982).

[Pen89] **Peng Y, Reggia JA.** A comfort measure for diagnostic problem solving. *Information Sciences* (1989) **47** *ii* 149-84.

[Phi91] **Phillips S, Wiles J, Schwartz S.** A comparison of three classification algorithms on the diagnosis of abdominal pains. *Proceedings of the 2nd Australian Conference on Neural Networks (ACNN 91)* (1991) 283-7.

[Rum86a] **Rumelhart DE, Hinton GE, Williams RJ.** Learning representations by back-propagating errors. *Nature* (1986) **323** *ix* 533-6.

[Rum86b] **Rumelhart D, McClelland J.** *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vols 1 and 2.* Bradford Books/MIT Press: Cambridge (1986)

[Rus83] **Russek E, Kronmal RA, Fisher LD.** The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis. *Computers and Biomedical Research* (1983) **16** *vi* 537-52.

[Sal91] **Salzberg S.** Distance metrics for instance-based learning. *Lecture Notes in Artificial Intelligence* (1991) **542** 399-408.

[Ser85] **Séroussi B, The ARC & AURC Cooperative Group.** Comparison of several discrimination methods: Application to the acute abdomen. *Lecture Notes in Medical Informatics* (1985) **28** 12-8.

[Ser86] **Séroussi B, The ARC & AURC Cooperative Group.** Computer-aided diagnosis of acute abdominal pain when taking into account interactions. *Methods of Information in Medicine* (1986) **25** *iv* 194-8.

[Sey90] **Seymour DG, Green M, Vaz FG.** Making better decisions: construction of clinical scoring systems by the Spiegelhalter–Knill-Jones approach. *British Medical Journal* (1990) **300** 223-6.

[Sho76] **Shortliffe EH.** *Computer-based medical consultations: MYCIN.* American Elsevier: New York (1976).

[Sho79] **Shortliffe EH, Buchanan BG, Feigenbaum EA.** Knowledge engineering for medical decision making: a review of computer-based clinical decision aids. *Proceedings of the IEEE* (1979) **67** *ix* 1207-24.

[Shw91] **Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF.** Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of Information in Medicine* (1991) **30** *iv* 241-55.

[Sim90] **Simpson PK.** *Artificial Neural Systems.* Pergamon Press: New York, Oxford etc. (1990).

[Spi84] **Spiegelhalter DJ, Knill-Jones RP.** Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society (Series A)* (1984) **147** *i* 35-77.

[Spi86] **Spiegelhalter DJ.** A statistical view of uncertainty in expert systems. In **Gale WA (ed).** *Artificial Intelligence and Statistics.* Addison-Wesley: Reading (1986) 17-55.

[Spi88] **Spivey JM.** *The Z Notation: A Reference Manual.* Prentice-Hall: Hemel Hempstead (1988).

[Sta92] **Stamper R, Todd BS, Macpherson PM.** A software engineering approach to the design of a medical expert system. *Proceedings of the 4th International Conference on Software Engineering and Knowledge Engineering (SEKE 92)* (1992) 341-8.

[Sta93] **Stamper R, Todd BS, Macpherson PM.** Case-based explanation for medical diagnostic programs, with an example in gynaecology. *Methods of Information in Medicine* (1993) (*in press*).

[Sto92] **Stonebridge PA, Freeland P, Rainey JB, Macleod DAD.** Audit of computer-aided diagnosis of abdominal pain in accident and emergency departments. *Archives of Emergency Medicine* (1992) **9** 271-3.

[Sut89a] **Sutton GC.** Computer-aided diagnosis: a review. *British Journal of Surgery* (1989) **76** *i* 82-5.

[Sut89b] **Sutton GC.** How accurate is computer-aided diagnosis? *Lancet* (1989) 905-8.

[Tel88] **Telfer S, Fenyö G, Holt PR, de Dombal FT.** Acute abdominal pain in patients over 50 years of age. *Scandinavian Journal of Gastroenterology (Supplements)* (1988) **144** 47-50.

[Tit81] **Titterington DM, Murray GD, Murray LS, Spiegelhalter DJ, Skene AM, Habbema JDF, Speke GJ.** Comparison of discrimination

techniqnes applied to a complex data set of head-injured patients. *Journal of the Royal Statistical Society (Series A)* (1981) **144** *ii* 145-75.

[Tod89] **Todeschini R.** *k*-nearest neighbour method: the influence of data transformations and metrics. *Chemometrics and Intelligent Laboratory Systems* (1989) **6** 213-20.

[Tod93a] **Todd BS, Stamper R.** Limits to diagnostic accuracy. *Medical Informatics* (1993) (*in press*).

[Tod93b] **Todd BS, Stamper R, Macpherson PM.** The design and construction of a medical simulation model. *Manuscript submitted to Computer Methods and Programs in Biomedicine* (1993).

[Wel89] **Wellwood JM, Spiegelhalter DJ.** Computers and the diagnosis of acute abdominal pain. *British Journal of Hospital Medicine* (1989) **41** 564-7.

[Wel92] **Wellwood JM, Johannessen S, Spiegelhalter DJ.** How does computer-aided diagnosis improve the management of acute abdominal pain? *Annals of the Royal College of Surgeons of England* (1992) **74** *i* 40-6.

[Wey75] **Weyl S, Fries J, Wiederhold G, Germano F.** A modular self-describing clinical database system. *Computers and Biomedical Research* (1975) **8** *iii* 279-93.

[Whi86] **Whitfield CR.** *Dewhurst's Textbook of Obstetrics and Gynaecology for Postgraduates.* 4th ed. Blackwell Scientific Publications: Oxford, London etc. (1986).

[Won90] **Felix Wong WS, Leung KS, So YT.** The recent development and evaluation of a medical expert system (ABVAB). *International Journal of Biomedical Computing* (1990) **25** *iii* 223-9.