The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Dr. Gordon John Brooks

Royal Free Hospital School of Medicine

A thesis Submitted in part fulfilment of requirements for the degree of Doctor of Philosophy in the Faculty of Medicine of the University of London

January 1993

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Abstract

Relevant literature has been reviewed regarding the performance, implementation and evaluation of computer based medical decision support systems.

The diagnostic performance of five simple case driven acute chest pain advice systems, have been compared using a standardized set of clinical records. A Bayesian inference model demonstrated superiority over two derived by logistic regression. Small data set flow charts performed well but both relied upon the use of expert opinion.

A Bayesian acute abdominal pain diagnostic advice system has been evaluated in a clinical trial. Standardized data collection improved the diagnostic performance of doctors. In practice, the computer system offered little additional user benefit. From further tests in primary care, it was concluded that, whereas general practitioners might enhance their performance by using data collection sheets, paramedics might benefit through direct use of the computer.

DERMIS is a new dermatology primary care diagnostic advice system. Components include a database derived from 5203 prospectively collected clinical records, a user interface, and an enhanced Bayesian inference model incorporating combined frequency estimates, expert beliefs and rationalized end-point groups. On laboratory testing, the diagnostic accuracy of DERMIS was 83%. The correct diagnosis appeared in the top three, of a possible 42 disease list on 97% of occasions.

In a semi-field trial of DERMIS involving 49 general practitioners, doctors did not always collect the same information as a dermatologist but were able to significantly increase their chance of making a correct diagnosis through use of the computer system. It has been concluded that although implementation of DERMIS might well increase general practitioner diagnostic accuracy and lead to improvements in the management of skin disease in primary care, rates of referral for specialist opinion might not be affected unless standard management plans are adopted.

Acknowledgements

I am most grateful to the following people and organisations;

Dr Richard Ashton, the Royal Navy consultant dermatologist for sharing his clinical expertise and allowing access to his clinics.

Doctors and staff of the surgical department at the Royal Naval Hospital Haslar for their assistance and participation in trials of acute abdominal pain advice systems.

Dr Paul Collinson, clinical pathologist, at the West Middlesex Hospital, for his assistance with the theory and application of diagnostic advice systems in cardiology.

Dr Roger Pethybridge, Head of Statistics at the Institute of Naval Medicine for his guidance in statistical matters.

Professor Tony Winder, of the Royal Free Hospital, London for his practical support.

The Ministry of Defence for giving me the opportunity to undertake this research.

My wife, Christine and the rest of the family whose boundless support and understanding made submission of this thesis possible.

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

List of Contents

Lists of Tables, Figures and Abbreviations

Tables

Tables (continued)

Tables (continued)

Figures

Abbreviations

AAP	Acute abdominal pain
ACP	Acute chest pain
AI	Artificial intelligence
AMI	Acute myocardial infarction
CAD	Coronary artery disease
CCU	Coronary care unit
CI	Conditional independence
CK	Creatine kinase
ECG	Electrocardiogram
GI	Gastrointestinal
IBM	International Business Machines
IBS	Irritable bowel syndrome
IHD	Ischaemic heart disease
MI	Myocardial infarction
NAASA	National abdominal pain study association
NSAP	Non surgical abdominal pain
NSE	Number of standard errors
OMGE	Organisation Mondiale de Gastroenterology
RL	Relative likelihood
RLQ	Right lower quadrant
RN	Royal Navy
RNH	Royal Navy Hospital
RNMSTS	Royal Navy Medical Staff Training School
USN	United States Navy
WHO	World Health Organisation

Publications are referred to in the text by use of bracketed numbers eg. (1).

Cross references within the text are made in square brackets and follow the system;

[Chapter.heading.sub-heading.item]	eg	[4.5.b.ii]
[Table number]		[Table 11]
[Figure Number]		[Figure 5]
[page]		[p 100]

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Introduction

The broad aims of this work have been to;

a) Study the design and evaluation of past and current diagnostic advice systems in at least three clinical domains, in order to identify adopted methodology and other factors that appear to make systems suitable for use in clinical practice.

b) To evaluate, through experimentation, the strengths and weaknesses of various diagnostic advice systems that have found use in clinical practice. To assess the suitability of such systems for the roles they were designed to fulfill.

c) To investigate, and where possible enhance, the function of simple inference mechanisms used in diagnostic advice systems.

d) To document and discuss the requirement for, design, construction and evaluation of a new medical diagnostic advice system.

The first aim has been met in part by the review of literature and is completed during later discussion. The other aims have been pursued through experimentation.

Three medical domains that encompass the causes acute chest pain, acute abdominal pain and skin disease have been investigated in depth. Advantages have accrued from studying each of the domains.

A number of advice systems that have been designed to assist casualty officers with the task of identifying patients suffering with 'high risk' chest pain have found use in clinical practice. They use differing methods of diagnostic inference and different lists of clinical variables.

Comparative testing has revealed strengths and weaknesses of the methods adopted.

In contrast, a single advice system designed to assist junior surgeons with the diagnosis of patients suffering with acute abdominal pain has probably been subjected to a more detailed programme of evaluation than any other diagnostic advice system. It has been shown to confer advantage upon its users and upon patients, yet it is not in widespread use. Here, the opportunity has been taken to conduct an independent field trial of the system in order to assess its suitability for practical clinical use. Investigation has also been performed into a possible primary care role for the system.

The development of DERMIS has been described in greatest detail. The work has involved the collection of a large database of over 5300 case records and investigations of the ways in which the intrinsic descriptions of disease locked within the case records can be represented in order to assist with the prediction of the presence of disease in new cases.

In the first two chapters of the thesis the history and evolution of medical decision support systems are put into the context of the clinical requirements of the three medical domains chosen for particular study.

The theme of the second chapter is medical advice system evaluation. This chapter closes with a detailed justification and planned order for the investigative work to be carried out. The ordering is maintained throughout the remaining methods, results and discussion in order to allow logical progression and easy cross reference between the sections.

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Chapter 1

Medical Decision Support Systems: Evolution and Medical Context

1. The Role of Diagnostic Advice Systems

Medicine, the art or science of prevention and cure of disease, is a vast, expanding and evolving subject. Medical practitioners cannot hope to keep up with all its developments and tend to concentrate on parts of the subject. In practice this has lead to the development of specialities and the recognition of specialist or expert practitioners.

In the United Kingdom, there is separation into primary and secondary medical care. Patients initially consult general practitioners who have broad but necessarily focused knowledge of medicine. A minority of cases cannot be dealt with by general practitioners and are referred to secondary care doctors for further guidance, investigation or specific treatment. Secondary care tends to be based in hospitals, but even here expertise is shared amongst generalists, specialists and sub-specialists.

In both the community and in hospitals, practical day to day patient care is often provided by nursing and paramedical staff who rely upon registered medical practitioners for support in clinical decision making.

This hierarchy of care workers seeks to provide health care to the community in an efficient and effective manner. However, the person most qualified to deal with a patient's problems may not always be the person who is first consulted. This is not unreasonable as a patient will often not know which expert to consult and in any case, more often than not, a specialist opinion will not be required.

Medical decision support systems, including those that provide diagnostic advice have been proposed as a means of enhancing the performance of generalists by giving them rapid access to the knowledge and decision making capabilities of specialists (1).

It has been argued that GP's are most in need of decision aids because they see a variety of problems at early stage, have restricted resources and few colleagues available for immediate consultation (2).

Medical diagnostic advice systems have appeared in many shapes and forms. Text books can be considered as being, on the whole, passive advice systems. They contain distillates of expert knowledge and experience but users must determine which parts are appropriate to current decision making. Active systems could assist with the recording of information and its evaluation, diagnosis and treatment (1)

The Three Clinical Domains of Study

Throughout this thesis particular attention has been paid to the application of diagnostic advice and other decision support tools to three clinical domains. These are the diseases that cause acute abdominal pain, the diseases that cause acute chest pain and diseases of the skin. The three domains will now be introduced with reference to diagnostic decision making and possible roles for advice systems.

2. Acute Abdominal Pain

a) The Clinical Importance of Acute Abdominal Pain

Abdominal pain is a common symptom that is associated with a variety of surgical and medical diseases whose consequences range from being trivial to life threatening. Abdominal pain of sudden onset will often cause a patient to urgently seek medical advice. Acute abdominal pain (AAP) was defined by de Dombal as having a duration of one week or less (3,4). An indication of the proportions of each of the common causes of AAP likely to be found amongst patients attending hospital casualty departments are shown in Table A.

The two most common causes of AAP are Appendicitis and Non-Surgical Abdominal Pain (NSAP). In the UK, approximately 55,000 cases of appendicitis and 120,000 cases of NSAP present to hospital each year (NASA). NSAP is a general term that refers to all causes of abdominal pain that do not require surgical intervention (5) including for example, mild gastro-enteritis and a urinary tract infection. Appendicitis is a surgical emergency for which the recommended management is laparotomy and excision of the appendix. If operation is delayed the appendix may perforate and its contents spill into the abdominal cavity causing generalised peritonitis which is a life threatening complication.

Conditions such as a peptic ulcer and diverticulitis also lead to perforation of the gut and peritonitis. Other organs such as the spleen, aorta, and a fallopian tube in ectopic pregnancy, can rupture or be ruptured and lead to an 'acute abdomen' that requires surgical intervention (OHCM).

The nature, severity and location of AAP varies between diseases. A classical case of appendicitis might present with acute abdominal pain that is initially central and colicky, but which moves after several hours to become steady and located in the right iliac fossa. Renal colic is a particularly severe form of AAP that is commonly associated with the blockage of a ureter by a stone and may require surgical intervention (6).

In this study I will be investigating the use of decision support in the management of patients suffering with AAP in three clinical settings that include the hospital environment, remote locations and general practice in the community.

b) Management of Acute Abdominal Pain in Hospitals and Decision Support

Patients who develop AAP might seek the advice of their general practitioner or be taken straight to the casualty department of a hospital. Depending upon his diagnosis, a general practitioner may offer treatment at home or arrange for hospitalisation. In the casualty department, the casualty officer may seek the advice of surgeons in deciding whether or not a patient should be admitted for possible operation. Once a patient has been admitted to a surgical ward, decisions need to be taken concerning the necessity for and urgency of operative intervention.

We have seen that in some cases of acute abdominal pain timely surgery and therefore hospitalisation is essential, whereas in others treatment may be conducted at home. The timing and nature of management offered can depend upon decisions taken by a patient, his general practitioner, the casualty officer, and junior and senior surgical staff. If this screening process were efficient then we might expect that there would be few unnecessary admissions to hospital and few inappropriate operations carried out. However, in practice it has been found in the UK, for example, that some 175 thousand cases of 'suspected appendicitis' are admitted annually of which a third are finally diagnosed as having had the disease. Of patients who are suspected of having appendicitis, some 16% are operated on but are found not to have suffered the disease. Approximately 23% of patients with appendicitis go on to suffer the complication of a perforated appendix.

The performance of hospital staff dealing with patients suffering with AAP has been extensively studied (3,4,7,5,8,9). When diagnostic accuracy was measured for various grades of surgical hospital doctor it was discovered that a house officer could be expected to produce the correct diagnosis on 50% of occasions, whereas senior house officers might attain an accuracy of 60%. Consultants were found to be able to perform at even higher levels of accuracy.

It was postulated that the use of a computer based advice system might improve the diagnostic accuracy of junior hospital doctors who are called to deal with patients suffering with acute abdominal pain. This in turn might promote more appropriate management decision making and lead to measurable improvements in health care (3,4).

c) Acute Abdominal Pain Management in Primary Care

(i) The Remote Location

The development of AAP whilst at sea is probably one of the most feared ailments of the seafarer. Hester (10) found that AAP was the most common surgical emergency occurring on submarine patrols in the United States Navy (USN). In the Royal Navy (RN), the vast majority of the seagoing population is aged under 40 (11).

In the UK, Appendicitis and NSAP account for 77% of cases of AAP in males aged between 15 and 40, who are sufficiently ill to seek guidance at a casualty department (NAASA). Within this group, NSAP is found to be the cause of the pain more than twice as frequently as appendicitis. Although there is no significant difference between the incidence of appendicitis in young adult males and females (NAASA), the hospital presentation rate of females suffering with NSAP is higher than that for males. In young adults, 4% of patients develop peritonitis in every 12 hour period that the disease remains untreated (NAASA).

In the RN, medical officers spend most time at sea immediately following their house officer training. However, RN warships rarely carry a medical officer and much of the medical support to the Fleet is provided by paramedics or designated non-medics working alone. The larger ships are provided with a basic medical library.

In military service, a large number of young persons can be isolated from hospital facilities for many months. Unless a surgical specialist is present, the preferred management of a patient presenting with appendicitis at sea is evacuation to the nearest surgical facility. Such an evacuation can prove to be a major logistic exercise with both operational and financial penalties. The type of management provided to a patient who develops AAP at sea will depend largely upon the ability of a ship's medical personnel to discriminate between appendicitis, NSAP and other causes of the symptom.

(ii) General Practice

A patient in the community who suffers AAP might well seek the advice of his general practitioner. The general practitioner has to decide whether immediate referral to hospital is appropriate. If he decides upon referral then care of the patient is delegated to doctors at the hospital. If, however, he concludes that referral is currently inappropriate then he must decide whether any further monitoring is necessary, how frequently this should be performed and the criteria that might indicate subsequent admission.

d) Decision Support for AAP Management in Primary Care

General practitioners working in the community and medical officers, paramedics and other military staff responsible for health care in remote locations do not normally receive specialist surgical training. It is suggested, therefore, that they might benefit from the use of a decision support system that increased their diagnostic accuracy and lead to improvements in the management of patients presenting to them with acute abdominal pain.

The USN faced with the same problem decided that computer based decision support might be employed to enhance the ability of paramedics who were required to make decisions concerning the evacuation of patients with acute abdominal pain (12,13). They adopted the Leeds acute abdominal pain advice system on the basis of reports of its diagnostic performance in hospital based studies (14,15).

The management decisions made about patients with acute chest pain have similarities to those made about patients with acute abdominal pain. Both symptoms can indicate the presence of life threatening disease and are common reasons for patients to urgently seek medical advice. In both cases general practitioners and casualty officers need to decide which patients require specialist medical care.

3. Ischaemic Heart Disease

a) Epidemiology

Ischaemic heart disease (IHD) is a common cause of acute chest pain. The regional prevalence rate of IHD in the UK amongst men aged between 40 and 59 of as measured by a WHO chest pain questionnaire and electrocardiogram (ECG) has been found to vary between 17% and 30% (16).
Transient ischaemic episodes with self limiting pain is described as angina, whereas prolonged pain with permanent ischaemic change is referred to as myocardial infarction. In the UK, IHD accounts for some 354 annual deaths per hundred thousand of the male population aged 35 and over (WHO figures, 1988). This represents one of the highest national IHD mortality rates. Although the IDH mortality rate for British women is much lower at 269 deaths per hundred thousand, this is still high in relation to equivalent international rates.

In some 'western' countries such as the United States, Australia, Canada and New Zealand, the mortality rate for IHD has fallen in recent years. However, in the UK there has been little change in the past decade (17).

b) Ischaemic Heart Disease: Symptoms and Signs

Frequently, the sudden onset of heart muscle ischaemic pain is the first indication of heart disease. Cardiac ischaemia probably develops when myocardial demands for oxygen exceed the capacity of the diseased vessels to supply blood. Coronary vasoconstriction may also increase resistance and thus reduce blood flow. Vasoconstriction can be mediated through neural pathways or caused by substances released by aggregating platelets and is believed to be associated with cigarette smoking, exposure to cold, exercise, endothelial injury (18) and possibly mental stress (19). Ischaemia is most common in the early morning (18) at a time when the fibrinolytic activity of blood at its lowest.

Only half of those with definite MI detected by ECG (16), however, are likely to have suffered chest pain suggestive of angina or MI. Epstein, for example, found that 70% of ischaemic episodes in patients with symptomatic coronary heart disease were not associated with angina and some 10-15% of acute MIs are silent (18).

Various other signs and symptoms may be present in acute presentations of IHD depending upon the site, extent and effect of any infarct and the physiological response to myocardial damage. Accordingly, the presentation of the disease in patients with ACP may vary from sudden collapse and death with pump failure due to arrhythmia to one of retrospective detection in apparent non-sufferers. Patients with acute symptoms suggestive of acute MI are often referred or present themselves to a hospital casualty department.

The likely presenting clinical features, results of investigations from, and risks to, a patient vary with duration of myocardial ischaemia. For example, in acute MI, the conductive abnormalities that cause the classic ECG features of Q waves, ST elevation and T wave inversion tend to develop during the illness, but an ECG taken in casualty soon after the start of pain may be normal. Similarly, enzyme indicators of heart muscle damage such as creatinine kinase (CK-MB) attain their peak serum levels many hours after the start of the illness.

c) Early Intervention

In recent years there has been renewed interest in definitive early treatment of acute MI (20). Current opinion is that decision making should begin in the first few minutes of arrival at hospital, as the early use of thrombolytics in patients with AMI has been shown to reduce the risk of death and help preserve the myocardium. Early intravenous streptokinase, for example, has been shown to result in recanalisation of coronary arteries in 55-75% patients (21,22). The greatest benefit has been demonstrated in patients with anterior infarction (23) where a 5% reduction in mortality rate is possible. The treatment is not without risk, however, and can only be given once, because of the likelihood of sensitisation. It is important, for this reason, that every effort is made to reach the correct diagnostic conclusion before the enzyme is given.

It has been recommended (20) that all patients who present within four hours of the onset of chest pain, without contraindications to treatment, should receive intravenous streptokinase with appropriate anti-allergic cover, but that consideration should still be given, however, to patients presenting within 12 hours.

The immediate treatment is usually followed up with further anticoagulent therapy such as short term heparin followed by daily aspirin. Revascularisation is possible at angioplasty, but is not always available. Patients presenting with chest pain and ECG changes (ST elevation or depression) are likely to benefit from intravenous nitrates which minimise phasic vasoconstriction. This treatment may also limit the extent of the infarct. Intravenous ß blockers may also reduce short term mortality if given to selected patients (24).

With time and investigation the correct diagnostic conclusion will be reached in almost all cases of acute CP.

However as the mortality of MI is greatest in the first 12 hours, with 50% occurring in the first two hours and 80% occurring in the first 4 hours of the onset of pain, it is also important that such a decision be reached quickly. If patients are admitted through the casualty department then a decision must be made by the staff as to when the patient will be seen by the doctor (20). Wyatt, for example, reported that patients with ACP might have to wait between 1/2 hour and 5 hours to be seen by the casualty officer (25). In assessing the patient, the casualty officer will have to decide whether the patient should be admitted or returned to the care of the general practitioner.

d) Hospital Admission

If a provisional diagnosis of MI has been made, initial management is normally carried out in a hospital coronary care unit (CCU), where facilities for expert monitoring and resuscitation are provided. Such units are costly to run and are best used if occupied only by high risk patients who have, for example, suffered acute MI.

The application of strict admission criteria may have the adverse effect of reducing the sensitivity of selection with persons at risk of developing complications being admitted to a general ward or being sent home. The management policy with the highest specificity would be to admit all patients with acute chest pain to the CCU. This would lead to unnecessary intensive management of some patients and possible exclusion of patients from the CCU when all the beds have been filled.

It has been reported that current practice often results in between 20% and 50% of CCU admissions being eventually diagnosed as not having MI (25, 26).There have also been reports that up to 15% of high risk patients are sent home

(27), rather than being admitted to the CCU.

As has been discussed, in the first few hours, ECG and enzyme investigations may be of no value in determining which patients require admission and intensive management (28). The diagnosis must be made in these cases on the history and clinical features alone. Unaided physicians have been shown to achieve specificity rates for AMI identification of between 66% and 93%. Poretsky (29) has suggested that physicians perform better when admission numbers are limited by CCU bed availability.

Diagnostic classification, in itself, is not necessarily a predictor of risk to the patient. In many cases, physicians assessing the necessity for CCU admission, try to identify high risk patients. These may include patients who have unstable angina.

e) Further Investigation

It is often the case that standard follow-up investigation of patients who have suffered ACP fails to prove or disprove MI and there is discharge from hospital without a firm diagnosis having been made. The clinical choice is to incur the expense of further investigation or to accept the provisional diagnosis. One of the further investigations that can be performed is coronary angiography, where the coronary arteries are viewed following the injection of contrast medium. In general, patients without coronary artery narrowing are unlikely to have suffered ischaemic pain and have a better prognosis than those where narrowing has been found. Coronary artery narrowing, however, does not necessarily indicate that patients have suffered ischaemia.

f) Non-Ischaemic Acute Chest Pain

Although ischaemic heart disease is the commonest serious cause of ACP others diseases such as pericarditis, aortic dissection, pulmonary embolism, pleurisy, nerve root lesions and chest wall pathology should not be overlooked when formulating a differential diagnosis. Gastro-intestinal diseases such as oesophagitis, peptic ulcer, pancreatitis and cholecystitis also possible causes of acute chest pain.

g) Role of A Diagnostic Advice System Advice System in Acute Chest Pain Mangement

The clinical problems in ACP can perhaps be summarised as follows;

(i) A patient with ACP may be suffering with life threatening IHD, where the mortality is greatest in the first 12 hours.

(ii) There are currently no simple tests that will guarantee confirmation of diagnosis within the first few hours of the onset of pain. At this time provisional diagnosis must normally be made on clinical grounds.

(iii) Patients presenting to casualty departments with ACP may have to wait in order to be assessed. Many acute ischaemic events are asymptomatic.

(iv) Treatment is available for AMI that is best given in the first few hours of the illness. The treatment itself is not without risk, but this can be justified providing that there is confidence in the diagnosis

(v) High risk patients including those with acute MI and unstable angina should be admitted to the coronary care unit for monitoring and further specialist treatment as required.

(vi) A decision concerning admission must be made within the first few hours of presentation. 15% of high risk patients are sent home. Up to 50% of patients admitted to the CCU are subsequently found not to be in the high risk group (25,30).

(vii) Incorrectly classified admissions should be given appropriate management once diagnosis is known.

Diagnostic advice systems could well be of value in the management of patients with ACP if they were able to influence the speed and accuracy of decision making.

4. An Advice System for Dermatology

a) Introduction

In considering the place for advice systems in dermatology, there is an immediate shift of emphasis from the high risk clinical decision making scenario of acute abdominal pain or acute chest pain to a subject where inappropriate decisions frequently mean little more than an extension of the required treatment time before problem resolution.

b) Skin Disease : The Clinical Setting

Skin disease is common in the community, but normally causes annoyance rather than a threat to life. For example, in 1987, the UK all age death rate per 100,000 of the population for dermatological disease was 2.9 as compared with that for ischaemic heart disease which was 628.3 (1987 Annual WHO statistics). For this reason, perhaps, lesions and rashes of the skin are often treated as being relatively trivial in nature, both by the sufferers and those that treat them. It is commonly, then, after some period of tolerance and perhaps self medication, that the problem causes sufficient annoyance or interruption of daily routine for a patient to seek medical guidance.

The diagnosis of skin lesions is in essence a problem of pattern recognition. A dermatologist can usually rapidly identify causative disease by visual inspection (31). Patients will, however, normally first consult a general practitioner for whom skin disease makes up 7% of all consultations (OPCS GP morbidity statistics). Many doctors find the accurate diagnosis of skin disease difficult which may be in part due to the limited amount of dermatology experience included within standard professional training programmes.

Immediate advice is available to primary care physicians in the form of text books and articles (31), packed with colour photographs. Unfortunately, although these may be useful in training they cannot normally be used during consultations. A search of the photographs and chapters that contained information about the observed features of a patient's condition could be time consuming, perhaps embarrassing, and would necessitate some knowledge of the possible diagnosis. The result is that patients suffering with even quite common skin disorders are referred to hospital clinics for advice concerning diagnosis and management. The management provided, rarely involves hospital admission or detailed investigation, and often takes the form of reassurance or a course of medication.

In the same way that many patients with acute abdominal pain are referred to hospital because appendicitis is suspected, a number of dermatology clinic referrals are made by general practitioners because of suspected malignancy. Common causes of concern are the pigmented lesion that might be a melanoma and the rough, expanding lesion that could be a carcinoma. General practitioners tend to adopt a policy of referring all such cases in an attempt to produce a high sensitivity in tumour identification, as it is well known that cure is possible with early treatment. If there is any doubt that the lesion is benign, the dermatologist will in turn commonly perform an excision biopsy. In many other cases, however, the patient will receive nothing but reassurance.

A case could be made for providing a dermatology diagnostic advice system in primary care if it could be shown that improvements in general practitioner diagnostic accuracy might lead to improvements in patient care.

5. The Use of Computers in Medical Practice

Calculating and computing machines can assist with data processing tasks. They offer powerful and reliable means of storing and rapidly manipulating vast quantities of numerical and textual information and will operate with the same efficiency despite time of day (1). However, they have no inherent common sense.

The widespread use of computers in all walks of life is a relatively new phenomenon. There is a necessary time lag between the development of new procedures that use computers and the dissemination of the methodology through teaching. Doctors who have had no basic computer awareness education have often felt uneasy about the adoption of automated techniques that might directly interfere with the practice of medicine as they know it (32). This situation is changing. In 1981, Teach and Shortliffe (33) detected the emergence of positive attitudes towards computers amongst doctors. In 1984, Kunz (34) noted that attitudes to computers in medicine were changing from initial scepticism through curiosity towards acceptance. As the years go by, greater proportions of medical graduates will be aware of the advantages and limitations of computer use.

In recent years, there has been a considerable increase in the numbers of computer based patient record and administration systems installed in both hospitals and general practice in the UK. This trend was initially stimulated in general practice by a commercially lead programme of offering a free system in exchange for the commitment to provide statistical information about patients. The UK government now offers general practitioners incentives in the form of re-embursement of installation and running costs to allow them to purchase systems of their choice. At the time of writing, nearly 75% of UK general practices have installed computer systems (35).

Numerous systems are available to general practitioners although the vast majority are supplied by a few companies. The UK government is promoting standardisation and partially funding trials of data exchange between general practice, health authority and hospital computer systems.

The UK government has sponsored the READ Clinical Classification system as a means of encoding clinical information held on health care computer systems. Other classifications commonly used in practice include ICD(9) and the Unified Medical Language System which is a superset of the majority of other known classification systems (36,37).

An advantage of adopting coding systems is the ability to store clinical and other information about patients in a way that can be easily retrieved and manipulated by computer systems. It can also facilitate communication with other computer systems and remote databases (37). Current coding systems have disadvantages. They are not as a rule issued with lists of definitions and are not always sufficiently rich to allow accurate representation of clinical findings (36).

Current general practice systems offer a wide range of facilities to their users including for example;
- the storage and retrieval of patient data
- a drug information and interaction database
- an appointments manager
- simple decision support tools such as template designers that allow for standardised data collection and protocol builders that allow users to construct and follow clinical protocols and management guidelines.

A straight forward application of computer technology to the problem of assisting with medical decision making has been the appearance of text books, such as the Oxford Text Book of Medicine, on electronic media. Flow charts can also be represented and computer programs written to guide the user through the embedded logic. Canon and Gardner (38), have found that compliance in the use of flowsheets can be increased by transferring them to a computer.

It is a short step from this position to the provision of interactive decision support (39). As a simple example, Ornstein et al. (40) reported a module of a computer system could assess whether a patient attending for consultation was in a population group whose members were being offered routine screening, for example of serum cholesterol in males with a high risk of heart disease. If so, the computer immediately generated and displayed a reminder for the doctor. The result was improved compliance with agreed screening programmes.

Another program has been developed to screen hospital inpatient records for evidence of adverse drug reactions. Over an 18 month period the program identified 731 new, and subsequently verified, cases among 36,635 admissions. Only 9 cases were detected by traditional methods (41).

Reminder and text book systems have also appeared on computer systems in hospital practice. For example, HOIS, the house officer information system was an electronic text book that contained data sheets of symptom causes and suggested managements (32).

6. Diagnostic Advice Systems: The Form and Usage of Simple Models

A model of the diagnostic reasoning process was set down by Ledley and Lusted in 1959 (42). Since that time many have been inspired to find algorithms that could be used by computers to actively assist in the process of medical decision making.

a) Clinical Algorithms

Clinical algorithms or protocols are flow sheets that are normally designed by expert clinicians with the aim of assisting less experienced doctors with diagnosis and management (34,1). They are essentially the expert's "descriptive" perceptions of his own reasoning process (43,44) and as such can form a link between the knowledge in a text and the ways in which an expert might use it.

Clinical algorithms are well accepted within the medical profession and are normally quick to use (32). They often cope well with common presentations and provide a basic plan upon which to base investigation and other action.
They cannot be expected to cope with all circumstances and

presentations, however, and the limitations can often be found by applying the details of a difficult case, when a course of action may not be found. Other common problems are continuous loops and ambiguous questions.

A number of clinical algorithms have been developed by Lynch (43), which are designed to enable the user to classify dermatological disease by stepwise analysis of the presenting features. The stated objective has been to provide a short differential list of likely diseases. Although simple and quick to use, the system has potential limitations. Only a small number of diseases are considered and there is minimal definition of the terms used. In common with other clinical algorithms, no evaluation is reported.

A more ambitious project has been undertaken by Ashton (44). He has produced a book of dermatology algorithms with supporting definitions and clinical photographs. New cases can be classified by following branches of the tree from page to page. A single diagnostic solution can be reached in each case. There has been no formal evaluation of the system.

Mc Donald (45), suggested that clinical protocols can help reduce the effects of information overload and thus improve quality of care. He found however that this effect only applied when clinicians had agreed the protocol before use and was reversed by removal of the decision aid (46). Smith (47) in a BMJ editorial bemoaned the general lack of agreement within the medical profession as to how management protocols should be produced or by whom and what bodies should sanction their use.

b) Statistical Models

Statistical methods have been employed in diagnostic decision support for many years (34,48,49) and have included discriminant analysis (50,51), linear regression (48,52,53,54), cluster analysis (48,55), pattern recognition (56), scoring systems (57,58) and Bayesian analysis. In many ways, the statistical methods of prediction form a coherent group which relies, to a large extent upon the benefits and limitations of averaging (59).

Bayes Formula

(After Thomas Bayes 1702-61):

If disease D (e.g. appendicitis) and sign S (e.g. rebound tenderness) are independent events then (60,61);

$$P\ (D|S) = \frac{P\ (S|D) \times P\ (D)}{P\ (S)}$$

where P (D|S) is the probability of the disease being present given the sign elicited

P (S|D) is the probability of the sign occurring in the disease

P (D) is the (prior) probability of the disease

P (S) is the probability of the sign occurring

Further details are given [in chapter 3]

Application of the Bayesian method thus takes into account all that is known about a new case and can be used to produce a relative prediction (43) or differential listing from amongst the included (exhaustive) set of diseases.

7. Bayesian Diagnostic Advice Systems

In terms of diagnostic advice system production, Bayesian analysis probably the most commonly employed 'statistical' method. Examples can be found for its use in jaundice (62), chest pain (63,64), cerebral disease (65,66), head injury (67), gastroenterology (15,50,68), upper GI bleeding (69), dentistry (203), vaginal discharge (70) and rheumatology (71).

Its popularity peaked in the early days of diagnostic advice systems but declined following the widespread adoption by research workers of 'expert system' techniques. In recent years there has been a revival of interest in the application of Bayesian methods in the construction of simple clinical diagnostic aids (71,72,73).

a) The Leeds Acute Abdominal Pain Advice System

One of the most well known clinical advice systems that uses Bayesian inference is the Leeds acute abdominal pain diagnostic advice system.

In 1972, de Dombal et al. (3) first described a computer program that could offer a diagnostic solution in cases of acute abdominal pain. They attracted considerable interest, and probably disbelief, by claiming that the program could diagnose new cases with an accuracy of greater than 90%, a performance that matched the best consultants and far exceeded that of junior surgeons. Further details of comparisons of accuracy were published in the early 1970's

(4) when it was suggested that junior doctors might be able to improve their performance through use of the program when dealing with patients suffering with AAP.

The Leeds abdominal pain system evolved from this work. Its basic structure has changed very little over the years and includes a single page multiple choice data collection sheet, a prospectively collected database of information derived from many thousands of hospital cases and program that uses a Bayesian algorithm to compare new case information with stored data in order to produce a suggested diagnosis. Various other teaching and explanatory modules have been included from time to time.

A clinician wishing to use the system is first required to select appropriate pieces of information about the patient's history and examination from a structured list of 132 elements on the data collection sheet. He or an assistant then enters the responses into the computer using a keyboard. The program performs a Bayesian calculation using frequency information from the database that relate to the chosen responses. As a result of the calculation, each of the considered diseases (between 7 and 13 depending upon the version) is assigned a relative likelihood score. The disease with the highest relative likelihood score is offered as the diagnosis. In order to justify and explain its decision, the system can select and display parts of a data bank of stored information about diagnosis and treatment.

b) Taking Disease Prevalence into Account

Prior probabilities are used in Bayesian analysis to take into account the prevalence of individual diseases. It can be argued that this is not appropriate when considering the prediction made in an individual case, which at some time must be of a member of a rare disease group (203). For this reason Ohmann (69) in a system for the diagnosis of upper GI bleeding set the prior probabilities of each of the four considered diseases to 0.25 (74).

On the other hand, the inclusion of prior probabilities will ensure that the most common disease will be chosen, if the system has reached no decision by the analysis of other data. In fact, this mimics the adage taught during medical training that

"Common diseases occur commonly"

a piece of advice that is designed to ensure that the junior doctor attributes appropriate importance to the members of his differential diagnosis set.

A differential diagnosis can be obtained by listing diseases in order according to their relative likelihood scores, with the highest scoring disease at the top. There may, however, not always be a clear favourite and minimum thresholds can be employed to prevent the a system predicting diseases with low relative likelihood scores (75). For example, a minimum relative likelihood for positive prediction of 85% was selected by Weiner in 1986 (203) for a dental adviser.

The balance of sensitivity and specificity can normally be selected, within the bounds of accuracy of the system by moving the decision boundary;

Moving the decision boundary for +ve identification of disease (d1)

```
                           Boundary
                            RL(d1)=
       RL(d1)..................|....................RL(d1)
       =0%                    50%                   =100%
              Increase  <------.----> Increase
              sensitivity            specificity
              Decrease                Decrease
              specificity            sensitivity
```

Setting upper and lower relative likelihood bounds for exclusion and assignment to disease d1

```
                 Boundary                Boundary
                  RL(d1)=                RL(d1)=
       RL(d1)........|.........|............|.......RL(d1)
       =0%          25%                    75%     =100%
                       <-------------------->
                          No decision made
                          in poor area of
                          discrimination
```

Threshold values are of particular value in selecting the sensitivity and specificity values necessary for screening. Davenport (76), arranged for organic disease screening to be conducted, amongst dyspeptic patients, by a non-medical interviewer using a Bayesian system. A prediction of risk category (high/medium/low) was produced such that the low risk group had a 10% chance of ulcer and 0.3% chance of cancer, whereas the high risk group had 20% chance of ulcer and a 10% chance of having cancer.

The value of identifying high risk groups is that priority can be given for conducting further, possibly invasive, investigation. Such screening might be cost effective in dyspepsia as the condition represents approximately 1% of a GP's workload and there is a high clinical false positive (referral) rate for organic disease (77).

c) Scoring Systems For Prediction

Although the computer has remained a central and 'glamorous' (78) element in the construction and operation of most diagnostic advice systems, there have been attempts to replace the complex calculations of, for example, Bayesian analysis by simple scoring systems that can be operated without a computer. Many such systems use a corruption of Bayesian method based use of weights of evidence (57,58) which are based on the measure;

$$\text{wt} = \frac{\text{sensitivity}}{(1\text{-specificity})}$$

for a particular feature with relation to a particular disease.

Logarithms of the weights can be summated and the totals for diseases compared in order to produce a differential ranking (79). Knill-Jones has suggested that adding up weights might be more acceptable to doctors than the use of 'black box' calculation (79). It has also been suggested the weights could be adjusted by according to perceived clinical importance (71).

In a comparison of such scoring systems with independent Bayesian analysis in rheumatological disorders and dyspepsia

(78), similar accuracy of prediction was obtained by both methods (71).

d) The Problems with Bayes

Teather (80) has commented to the effect that the majority of diagnostic advice systems that use statistical methods have relied upon unrealistic assumptions concerning basic distributions within the population. Diseases are often difficult to separate, for example intermediate states exist between Crohns disease and Ulcerative colitis that could acceptably be attributed to either group.

It has been suggested by Teather (81) and Feinstein (82) that as Bayesian analysis fails to exploit the redundancies and correlations within diagnostic information, misleading conclusions may result. This failure becomes even more important when it is realised that in developing a Bayesian system, there has commonly been a meticulous, time consuming and costly collection of the data (83). Additionally, once the algorithm has been applied it is often discovered that the relative probability statements made, fail to be reliably associated with true probability of disease or outcome predicted (83).

Szolovitz & Pauker (75) pointed out that a fixed Bayesian system would not allow for changes in the incidence of a disease or of its symptoms with time or location. There might also be problems in identifying the effects of coexistent pathology or drug treatment (34), a problem that is not confined to the use of statistical systems.

e) Non-Independence of Variables and Bayes

In considering possible reasons for failure in a Bayesian system, one of the first factors to consider is the validity of the assumptions made (80). Creators of Bayesian advice systems have often made broad assumptions about the relationships between clinical findings and disease states. Spiegelhalter has coined the expression 'idiots Bayes' to describe the resulting methodology.

One assumption has been that an exhaustive set of mutually exclusive diseases has been used (80,3). Ohmann when failing to produce satisfactory performance from a model for GI bleeding, attributed the difficulties of diagnosis based on clinical features to various interactions, including the presence of physiological states (69) within disease groups.

A second assumption of conditional independence (CI) has often been made and challenged. de Dombal (84), and others (50), have assumed that symptoms and signs found in patients with acute abdominal pain have been independent indicators of disease presence. Kronmal has suggested that although the CI assumption may not be wholly applicable to Bayesian and other statistical diagnostic systems, the effects of any dependence can be ignored (85). Séroussi (50), has suggested that CI applies to rare disease groups where associations have been lost through combination of disparate cases.
An example of the possible deleterious effects upon system performance, of assuming CI, has been described by Teather (80). He considered the two diseases (d1 and d2) which had two symptoms (s1 and s2) [Figure 1];

Figure 1
Effects of Feature Association

		From cases, actual first order distribution				Frequencies assuming CI		
Binary state	s1/s2	00	01	10	11	10	01	11
% Cases	d=1	50	0	0	50	50	50	50*50
% Cases	d=2	0	50	50	0	50	50	50*50

In disease 1, both symptoms were either present or absent. The occurrence of only one symptom was wholly associated with disease 2. When CI is applied and frequencies generated then this perfect discrimination is lost and (for this example) a Bayesian system would predict both diseases as being equally likely whatever the status of the two symptoms.

Hilden has discussed the robustness of the CI Bayes model and demonstrated a relaxed model where the effects of dependence upon outcome were minimised (59). He concluded however that on most occasions the CI model could be satisfactorily applied especially where there was the possibility of missing data. He suggested that in practice medicine tended towards CI for the following reasons;

(i) Where variables are known to be related they are often replaced by a single variable.

(ii) Where clustering is found within diseases, separate syndromes or categories are described.

c) The most widely used estimate of accuracy is frequency of correct classification. Violation of CI by dependency is likely to produce extreme odds but without any change in the relative order of the predictions.

Fryback (86),considered that any effects of dependence could be minimised by concentrating upon a reduced number of variables.

Spiegelhalter and knill-Jones suggested weights of evidence using in scoring systems could be adjusted by logistic regression in order to take non-independence of disease features into account (79,87).

f) Minimum Data Sets

One method of minimising the chance of dependence between variables is to reduce the total number of variables used by picking out those that are most important (80).

Stepwise variable selection has been used in a Bayesian system designed to assist in the diagnosis of upper gastro-intestinal (GI) bleeding (74,88). A computer intensive iterative procedure was adopted using the 46 variables available. Following independent frequency estimation variables were randomly added and the combination producing the best diagnostic result kept for the next stage. Addition continued until the diagnostic accuracy peaked, as determined using the non-parametric Spearman rank correlation coefficient (86), in this case with a feature set of 17. It was found that once the peak had been reached,the addition of further variables tended to slightly reduce diagnostic accuracy.

Teather used an iterative process of data set reduction to produce a decision tree for the differentiation of surgical from non-surgical jaundice (80). He started with 64 features and selected the feature that had the highest predictive value for surgical jaundice. He then partitioned the case database according the presence or absence of the feature. The confidence interval for the diagnostic accuracy of the tree was calculated. The process was repeated within the partitioned groups using further variables until the diagnostic accuracy peaked. The result was a minimal dataset which used the dependent combinations of the top three features (age, if transferred from another unit, and drug exposure). The reported accuracy of the tree was 81%. The system was no more accurate, however, than one that used CI Bayes on all the available features (80).

When studying cerebral disease Morton (83) determined that an independence model did not produce an appropriate representation of disease states. As a solution a diagnostic tree was constructed and found to improve classification. Between 1982 (89) and 1988 (90), Goldman used a partitioning method to develop a minimum data set clinical algorithm for identification of patients with ACP who had suffered MI and found that it produced more accurate classification of cases than a Bayesian model produced by de Dombal (91).

Ohmann tested a similar method of data set reduction under a variety of conditions by various splits of a case database and concluded that a single selected sequence of features may not be appropriate for other that the conditions under which it was set up (74).

If a purely dependent system, like a diagnostic tree, is set up then problems arise when subsets become small. Where CI is applied, intra and inter disease feature dependence may cause errors of prediction. Methods have been proposed, that allow dependence to be taken into account, within the framework of Bayesian method. One such method that has been applied is the Lancaster model (92) which utilises all first order feature combinations.

This was put to a practical test in AP (50) against surgeons and a CI Bayes model where allocation to the correct diagnostic group was the criterion for success. In the first test where allocation was to individual disease group, the surgeons and the CI Bayesian model performed equivalently at approximately 63% accuracy (note: this is not the Leeds AP system, but uses the same collection sheet), and the Lancaster variation produced 67.7% accuracy. In a second prediction of surgical vs non-surgical outcome, an accuracy of 70.8% was reported for the surgeons and 88.7% for the Lancaster variation of the Bayes model.

There are two reasons for adopting CI; to reduce the estimation of $P(S|d)$ to a reasonable calculation and to avoid the errors of estimate that would be involved with attempted calculation of the complete (D,S) array.

By considering all first order combinations (Lancaster), a compromise is reached whereby, the frequency estimates are still being made on fairly large groups but an element of inter feature dependency is taken into account.

8. Expert Systems as Thinking Machines

In problem solving, computers were perhaps seen initially as efficient calculating machines. However, with the evolution of the concept of computer as a "thinking machine" came new hopes for a medical expert role.

The view that clinicians see diagnosis as probabilistic was criticised as being over-optimistic (69,93,94,95) and it was suggested that as human decision making depends upon the manipulation of large amounts of symbolic knowledge, rather than any statistical method, computer decision aids should mimic this process (96).

More natural human skills were considered to include integration, judgement and hypothesis testing (97). It was pointed out that integration might be modified by experience, intuition hunches and response to cues (72). Eddy and Clanton suggested that as clinical problems become more complex, physicians move from algorithmic to heuristic reasoning (98).

There followed a shift in research emphasis from purely providing something that would work to detailed studies of the mechanisms involved in creation and implementation expert systems. This required research in such areas as knowledge representation, heuristic search, the use of natural language and models of the thought process.

A generally held outline model of the human diagnostic process was set down by Young (32) and followed the line;
- Collect appropriate information
- Generate hypothesis
- Test sub-goals
- If successful adopt diagnosis

where the first three steps are repeated as required.

Other aspects of the diagnostic process that are thought to have some effect on accuracy are the differential processing of concrete and abstract information, frequency of opinion revision, the emphasis placed upon positive findings and their perceived strength and the recognition of apparent pathognomic features (99).

Gorry (100), argued along the following lines, his support for the use of symbolic reasoning in medical advice (expert) systems;

- Clinical judgement is not based on a detailed knowledge of pathophysiology but on gross chunks of knowledge and experience from which rules of thumb are derived.

- Clinicians know facts but their knowledge is largely judgmental. Their derived rules allow them to focus attention and generate hypotheses quickly without a detailed search of all the information available to them.

- Clinicians recognise levels of belief or certainty associated with rules but do not use them in any formal statistical manner.

- It is easier for experts to state their rules in response to perceived misconceptions in others than it is for them to generate such decision criteria a priori.

It was envisaged that medical expert systems could serve as consultants for human decision makers, but must be employed in an area where experts out perform generalists (72) and less experienced specialists. They would use general and domain specific knowledge and employ symbolic reasoning and heuristics to infer what was not explicitly described. They would be able to construct and test hypotheses by using a systematic yet transparent methodology that mimicked the mechanisms employed by consultants (34).

"...trend from techniques based on general knowledge, such as statistical methods, towards emphasis on techniques based on domain-specific symbolic knowledge, such as diagnostic inference rules. This trend has resulted in an emphasis on systems which interpret and explain the clinical significance of their findings, rather than simply produce another number for the user to interpret."

(Kunz, 1984)(34)

9. Construction of Expert Systems

a. Methods and History

Expert systems have classically been described as having three main modules, the knowledge base, a control structure (inference engine or interpreter) and an interactive man-machine interface. The knowledge base has commonly been derived by a knowledge engineer whose task has been to elicit information and heuristics from an expert and represent them in a logically coherent form that can be used by a computer.

This has often required the use of a specialist computer language such as LISP or PROLOG. The use of formal logic to describe disease was not a new concept, however. In 1955, Dale (101), proposed the use of Boolean algebra in psychiatry. Ledley & Lusted (42,102), went on to described disease and their symptom sets as Boolean functions. Feinstein (103) and Wulff (104), employed Venn diagrams. In 1981, Burton (105) proposed the use of formal logic in dermatology (for acne) where definition is often "imprecise".

By the early 1980s several medical expert systems had been developed and reported upon in the literature (106,107,108,109).

Shortliffe (110), described MYCIN which is a therapy adviser for infectious diseases. It employed some 400 (if... then) production rules as descriptions of the clinicians knowledge. Flow of operation in pursuit of sub goals was controlled by meta rules. Backward chaining was used to call up sub goals and specific questions that could be used to support the current hypothesis. In practice it was possible for the same question to be asked several times in pursuit of similar sub goals. The rules were written in such a way that they could be readily understood by clinician and computer alike and were displayed in support of offered therapy suggestions.

Another well known expert system, INTERNIST -1 (49) was designed to produce advice on test selection and diagnosis in internal medicine. Miller (49), considered that its reasoning process was unduly restricted by not being able to refer to anatomical and pathophysiological (deep) information that might be important in diagnosis. He observed that;

"The issue is whether such artificial intelligence models can reach conclusions similar to those of a competent clinician and can justify those conclusions in a rational and clinically acceptable way"

Details emerged of the construction and usage of other early systems which include ONCOCIN (111) which was constructed to assist with the management of patients on chemotherapy protocols and PUFF (112) which interprets pulmonary function data that is derived from a lab computer.

b) Expert Opinion

AI workers appeared to assume that medical experts are correct and with this justification used their knowledge for prediction (113). Carroll (72), discussed doctors perception of their own skills and pointed out that clinicians are often not aware of their own fallibility and can become overconfident, especially with experience. This confidence estimate does not necessarily equate with the difficulty of the case, however (32). One way for doctors to establish personal confidence limits is to monitor results of recommended treatments. Unfortunately, it is human nature to be selective in this process and to rationalise poor decisions. The evidence available to clinicians monitoring their own performance might also be biased in that patients affected by poor decision making might not be heard from again (32).

Even when clinicians identify the correct diagnosis it has been suggested that they might not always provide optimal treatment and even allow harm because they are wary of side effects (72). The choice of treatment is also in part determined by standard practice which may not represent the best option for the patient (99).

Some have expressed doubts about methods of expert selection (72). In INTERNIST, the expert simply volunteered for the task of providing knowledge for the system. It seemed that the expertise could not always be elicited and in such cases there was evidence that experts might be forced into producing explanations that they did not necessarily use themselves (72).

> "But who is to say what experts make the rules? Experts to some, to others may be fools!"
>
> (de Dombal,1983)

Many authors (34) have reported that although the early expert systems appeared to mimic the reasoning styles of experts (96,98), they had drawbacks in that they tended to take up a large amount of computer space, were slow and did not adequately represent the experts knowledge (114,115). It seemed possible to produce general rules for easy cases but in more difficult cases there was incomplete knowledge available and the addition of rules that detailed the exceptions and qualifications could rapidly make a system unmanageable (72).

c) Expert Systems, Difficult Cases and Deep Knowledge

It has been recognised that difficult diagnostic problems are sometimes caused when patients present with more than one condition (116) or the presentation of a condition changes with time. A rule based expert system might well fail on these occasions by trying to fit all the presenting symptoms and signs to one pattern.

In TNET, a temporal network extension to ONCOCIN (117), symptoms, signs and test results are time stamped and are assigned a life span of activity, according to their nature, during which they can be used in inference.

A logical method of dealing with possibility of the co-occurrence of two diseases in the same patient was proposed by Reggia (118), in 1985. The basic principles were that, if a manifestation can occur in a disease then this is a reason for the disease to exist in the differential. If, on the other hand, it cannot be explained by one disease then this is a reason to postulate the occurrence of two diseases simultaneously.

Non-numeric probability terms were included as a weighting scheme to order the differential and provide detail to the justification.

Reggia (119), developed 'set covering theory' to include a network of causal associations, hypothesis and test algorithms. The system, NEUREX, was set to accept the most simple solution (parsimony) to any clinical problem (providing that there had been minimal error in disease description and data collection). The stated advantage of the system was its 'deep' nature, whereby, it was able to recognise that perhaps both a disease and effects were present (119). For example, in MI, cardiovascular shock might provide confounding variables. It is perhaps worthy of note that, in this case, the poor condition of the patient might also reduce the amount of data that can be collected (69).

Wu (116) criticised Reggia's problem decomposition method for imposing an artificial structure on medical problems by the minimum list of candidates that parsimoniously explain a set of symptoms. He suggested that multiple disorders can be separated by symptom decomposition (116), by finding explanations for coherent groups of symptoms and signs (120,121). This could be incorporated into a structured approach to decision making by looking for common themes to get a minimal set of possible solutions that was clinically intuitive, for example; the presence of heart disease in a patient with hypertension.

He argued that it would be appropriate for an advice system to offer intermediate hypotheses as it could assist practitioners to
interpret symptoms, signs and lab tests, discard 'red herrings' and determine causal and temporal relationships.

Warner (122), suggested that no one model was appropriate for all applications and described the HELP system which contained a mix of clinical algorithms, decision analysis procedures, mathematical and statistical models and was integrated with a computer based hospital record system (123).

Kunz (34), described AI/MM which is a physiological model of fluid and electrolyte balance which incorporates symbolic knowledge of anatomy and physiological function along with mathematical descriptions of the principles of physics and physiology.

Despite the potential sources of error described, the combination of Bayesian probability and formal logic perhaps provides a means of overcoming many of the limitations of systems that exclusively use derivations of one of the two methods (50).

d) Dealing with Uncertainty

In recent years, there has been an increasing support for the view that medical advice systems require a means of coping with both expert core knowledge that is largely dependent and the 'fuzzy' areas of uncertainty (39) that abound in clinical medicine. Many have favoured probabilistic solutions, but others have pointed out that the problem remains qualitative rather than quantitative in that we should be determining how to reason in the face of uncertain beliefs and findings rather than trying to attach numerical values to certainty (124).

One proposed solution has been the use of fuzzy set theory (125) which attempts to assign probability values to soft 'expert' opinion by classifying such terms as 'low','high', 'suspected', 'possible' etc which are commonplace elements of specialist medical language. The objective was to use a combination of definitive rules and probabilistic solutions to maintain the 'rich nature of medical thinking' (125).

Several informal numerical systems have been adopted in Expert system construction, for example certainty factors in MYCIN (110) and frequency and evoking weights in INTERNIST (49). Hajek (126) attached degrees of certainty to rules allowing weighted propagation of evidence.
The practical value of such methods rely upon the accuracy of clinicians estimates of likelihood. Leaper (127) found, for example, that clinicians often only produced poor estimates of the association between relevant symptoms and disease in acute abdominal pain.

'Bootstrapping' or feedback of cases into the system to alter false assumptions has been used in an attempt to overcome the problem of inaccurate initial setting of likelihood values and assumptions embedded in rules (128).

QMR (Quick Medical Reference) is a modern descendant of INTERNIST that holds profiles on over 600 diseases (129). The controllers have recognised the problem of calibrating the weightings of new profiles that have originated from authors working at different sites. Provisional profiles are created with reference to both recognised experts and available literature (commonly over 100 articles) in order to achieve a balanced representation. These profiles are then carefully scrutinised by an expert review panel before being accepted for use in the system.
Evidence has to be produced for each link between feature and disease and between diseases. An automatic frequency generator converts standard phrases into a numerical representation (129).

ILIAD is another large domain medical advice system that can offer advice on more than 950 diseases. It is basically a Bayesian system that uses a statistical database derived from the records of patients admitted to the University of

Utah hospitals and produces an output of diseases ranked by posterior probability. Where deficient, the database has been supplemented by estimates of probability provided by experts or extracted from the literature (122,130). There are also control structures and knowledge frames that allow the system to adopt a hypothetico-deductive approach to inference. In the current version, for example, if some of the features of a classical case of appendicitis are entered, the system will set up a hypothesis that appendicitis is present and try to elicit further evidence to support the diagnosis. If at any stage the rare but appropriate finding of left lower quadrant pain is entered, the system excludes the hypothesis irrespective of any further evidence for appendicitis that might accumulate (personal evaluation of ILIAD Nov 1992).

e) Causal or Belief Networks and Uncertainty

One way of describing a disease process in depth, in a coherent fashion, is to chart all of the known or believed interactions between causative factors, intermediate states and observable effects produced in patients (131,132). Such causal or belief networks have been incorporated into expert systems to provide frameworks for inference (133,134,135).

Lauritzen and Spiegelhalter (136) described a causal network to express clinical knowledge derived from experts in a graphical form in the MUNIN system. The graph described qualitative dependency in EMG (Electromyography) investigation, specifically the pathophysiological relationships between disease states and test results.

In the network a structure comprising 25 nodes linked nodes represented each muscle. Paths lead from the nodes representing disease states, through intermediate nodes to nodes representing 15 EMG findings.

Lauritzen and Spiegelhalter (136) went on to describe a probabilistic inference method based on Markov chaining (137) within a Bayesian framework, that allowed estimation of the likelihood of the outcomes resulting from alteration of the states of input nodes.

The prior probabilities of the disease states were used to set the disease nodes. The possible states of each node were described in conditional probability tables. Child nodes were dependent upon parent nodes such that the probability of a child could be calculated if parent nodes were known. It was assumed that joint probability for the structure equalled the product for a particular set of 25 states equals the product over the entries in the 25 conditional probability tables that feature the appropriate states.

The probability values assigned to the various nodes could be updated as evidence arrived at the 'findings' nodes. In order to allow flexibility in updating a undirected graph was created by marrying parents and triangulation, a process where nodes were linked in groups of three.

These self-sufficient 'cliques' of nodes were connected by 'separators' that allowed propagation of evidence through the graph in an ordered way.

Once evidence had been absorbed from nodes, they could be eliminated from the graph reducing its complexity. The structure was a coherent and comprehensive model and allowed conditioning by new evidence, hypothesis testing and investigation of new findings.

The method has the potential for coping with missing information in sparse arrays, providing an adequate causal network can be defined and local relationships are known.

Beinlich et al. (138) compared the use of this method with a message passing algorithm proposed by Pearl (135,138) in implementation of the ALARM automatic patient monitoring system for anaesthesia.

A graphical structure of 8 diagnoses, 13 intermediate variables and 16 findings was set up, where all nodes were assigned a set of exclusive and exhaustive conditional probability values for possible states.

They found that Pearl's network algorithm was hampered by a need to recalculate probability values after the arrival of each item of evidence, whereas the Lauritzen and Spiegelhalter algorithm could cope with multiple simultaneous inputs.

In practice, the ALARM system has to be able to offer a rapid analysis of the state of anaesthesia. Whereas Pearl took 8 minutes to update on the arrival of a standard set of 8 findings, the L&S model running on this same hardware took only 3 seconds, which was considered to be clinically acceptable.

Both the MUNIN and ALARM networks described a small number of well defined interactions. It has been suggested that if probabilistic inference algorithms were applied to large belief networks, the calculations involved would be too complex and (computer) time consuming for any real time use of the system. A probabilistic inference algorithm applied to belief networks in QMR (version QMR-DT), for example, averaged 94 minutes per consultation calculation in laboratory testing, whereas the standard system took less than a minute (139) per case on the same test set.

Solutions to this problem have been sought in the development of approximation algorithms (140,141) and precomputation of the most common pathways within the network. For example, Herskovits and Cooper (142) estimated that pre-calculation of 841 of the 98,304 pathways through the the ALARM network would allow the system to supply an immediate solution to new cases on more than 50% of occasions.

Potential sources of error in all systems that use belief networks are the accuracy of interpretation of any expert's beliefs and the accuracy with which the beliefs reflect disease processes (143). These errors might be minimised by building graphical models directly from clinical data (144) or by using errors to adjust beliefs (136,145).

It has been suggested that even if belief networks could be constructed for systems that covered a large area of medicine there might well be insufficient detailed statistical evidence available to allow the construction of nodal conditional probability tables (73,124). In addition, the statistical information which is available can vary in quality and applicability (124). In the Oxford System of Medicine (146), which has been designed to provide clinical decision support to general practitioners prior and conditional probability statements are included where known. An example is:-

> the conditional probability of weight loss given cancer is 0.4 (124)

The statistic appears to be of limited clinical value as it gives no information about the population from which it was derived or to which types of cancer it refers.
In practice the Oxford System Makes little use of such information in inference and reverts to pure predicate logic where quantitative information is sparse.

10. Iterative Approaches to Diagnostic Inference

Genetic algorithms and neural networks have been used in the production of diagnostic advice systems and are examples of models that can be 'trained' directly from clinical case information. Such systems are designed to learn by their mistakes and do not require belief networks or rules to guide their inference mechanisms.

Genetic algorithms model the genetic selection process (147). Categorical data such as the presence or absence of symptoms or signs are represented in 'genes'. Combinations of genes are formed into 'genomes' and the genomes used to predict the presence or absence of diseases. Genes are selected for inclusion in genomes according to a probability matrix that initially holds random values.
A set of cases, patient records (51) for example, where the relationships between input and output states are known, is used to train the model. There is iterative adjustment of the probability matrix according to the efficiency that genomes predict diseases in order to favour successful patterns. In this way a group of patterns of say symptoms and signs can selected as predictors of disease presence (148).

11. A Computer Brain

Neural networks have been designed to mimic the function of interconnecting nerve cells. A neural network is a multi-layered matrix of algebraic equations (51) which can accept input data and calculate an output based upon this information and the experience of past cases. The network is composed of layers of nodes or perceptrons. Evidence is entered to 'input nodes' which pass their output to one or more layers of 'hidden nodes' which in turn excite 'output nodes'. An input node may represent a clinical feature which is set to be present or absent. An output node may represent a disease with the state of the node indicating its likely presence or absence. 'Hidden nodes' are used in calculation but do not represent any clinical state. Nodes can excite or inhibit those to which they are connected.

Random weights are assigned to nodes at initiation. Node excitation is calculated using a logistic function and is related to the product of input values from evidence or previous nodes and assigned weights (149).

Case information is used to train the model and there is back propagation of errors of prediction leading to adjustment of the node weightings. The training set is repeatedly presented to the network until a stable prediction error rate is reached (150).

The number of input nodes, output nodes and layers of hidden nodes can be varied to alter the predictive performance of the network.

A neural network trained to recognise the presence or absence of myocardial infarction in patients with anterior chest pain has been described by Baxt (149). The 20 possible inputs were cardiovascular symptoms, signs and test results which included ECG abnormalities. The rest of the network comprised 2 layers of 10 hidden nodes and a single output node. Myocardial infarction was predicted by the system if the value obtained from the output node exceeded a predetermined threshold value.

The network was trained on 356 cases, but stability was only reached after exclusion of 5 'atypical' cases. The predictive accuracy of the network was subsequently tested on 331 further cases and compared with and found to exceed that of doctors who treated the cases.

Threshold values for decision making may be used to tune the system for appropriate sensitivity and sensitivity. For example, in another neural network trained to detect cases of MI the output node produced results between 0 and 1. When there was a low value (less than .2) the system predicted no MI, a medium value (.2 to .9) indicated uncertainty and a high value (greater than .9) was taken to predict the presence of MI (151)

A common criticism of neural networks has been that it is not possible to observe, and therefore check, the mechanism by which conclusions are being made (51,143). In addition, in contrast to the findings of Baxt, the reported accuracy of prediction in medical practice has generally not been any better (51) and on occasions worse than that of doctors that systems have been designed to assist (152,153,154,155).

12. Comparisons of Models

So far we have looked at individual models and their applications, how can we tell which model to use if there are no comparisons? Many of the choices made have been on theoretical grounds more and more complex models produced (143). Which is best?

Lucas (51) has compared the diagnostic performance of three case driven models, a back-propogation perceptron network, a genetic algorithm and discriminant analysis using clinical data about patients who were suspected of having gall stones in their common bile ducts. He split the 174 worked up cases between training and test sets and applied the models. For each case, 12 features and the final diagnosis were known. The training cases were used for iterative training of the genetic algorithm and neural network and for the production a linear discriminant function. In each case a single prediction was required for the presence or absence of stones in the common bile duct. When applied to the test cases he found no significant difference in predictive ability between the three models.

Hart and Wyatt (143) compared predictive accuracy of three and four layered back-propagation neural networks with a simple Bayesian algorithm using 174 training and 73 new records from patients who had presented to a casualty department with acute chest pain. In each case the final outcome was known and the patients had been classified as having a high, intermediate or low risk of serious complications. They found that the three layered network correctly classified more cases than the four layered network, obtaining an overall accuracy of 70%. The accuracy of the Bayesian model was slightly greater than that of the best configuration of the network and produced a cleaner separation of the disease groups which allowed easier tuning for desired sensitivity.

13. Decision Theory and Advice Systems

Decision theory considers the assignment of values to choices, such that the utility of a particular course of action can be calculated in terms of likely outcome (156,157,158).It has been described by Wigerz as the "systematic approach for arriving at optimal strategy" (39).

There is obvious potential for use of the technique in medical decision making. A large variety of tests and treatments are available, many of which partially duplicate value of others. The effects of biological variability must be considered along with invasiveness, risk and cost (159). The aim might be to produce an iterative work-up, planned in stages, where the costs and risks were commensurate with the clinical problem.

The adjunctive use of decision theory with any validated diagnostic advice system is appealing and might enhance the overall value of the system. It could be used to take into account the balance of sensitivity vs. specificity (160) of a particular test or advice system and help in the determination of whether further investigation would be likely to produce benefit (75,159,161). Greenes has described the system CASPER (159) which is a work up tool for decision support in clinical problems. It is reported to use prior probability and sensitivity /specificity data to assist in the selection of procedures and evaluation of results. Young described CARE (32) an automated decision support textbook for critical care that was designed to save life in trauma patients.

The main problem with the application of decision theory to medicine is the assignment of value to outcome. Conflicting inter expert and inter patient opinions make some numerical assignments virtually impossible (34). For example, what risk of death (34) from operation can be equated with a certain level of disability if no operation is performed? Wigertz suggested that success might depend upon physicians' willingness to formulate quantitative statements concerning incidence (39), prevalence and outcome, which Weinstein and Fineberg (156) described as

"an articulation of common sense"

The answers obtained might depend upon the surgeon concerned or a particular patient's tolerance of pain.

In the large domain medical decision support system QMR-DT, a belief network using statistical inference feeds into a utility model in order to produce advice on the cost-effectiveness of proposed investigations (139).

14. Explanation and Justification

In 1973, Gorry (100) postulated that the ability of diagnostic advice systems to explain inferences is of central importance to acceptance by physicians. Teach and Shortliffe (33) supported this, reasoning that physicians will hesitate to use such systems unless they can confirm the basis for advice given. Reggia (162), reasoned that systems designed to educate would need to provide answer justification, in order to meet that objective.

Reggia, made positive suggestions concerning features that were likely to be available in systems which could be used in justification. In Bayesian systems, analysis of the prior and conditional probabilities can produce understandable and plausible explanations. For rule based expert systems, the route to any conclusion and the knowledge statements used can be directly revealed as explanation (162). Where the method of construction of a system does not lend itself to providing explanation, a pre-formulated solution is sometimes offered (14,64).

Morton (83), considered that the most important features of a system's output should be an indication of most likely disease, any alternatives (50) and a measure of the certainty of prediction. Ohmann (69) criticised the use of posterior probabilities in explanation as they often do not represent 'true' probabilities occurrence and might mislead clinicians (163,94).

Miller has criticised the 'Greek Oracle' model of decision support where a physician supplies the computer with large amounts of information in return for a solution without any explanation (164).

In a similar vein, Kunz (34) directed an attack against simple probabilistic systems that normally have no access to explicit knowledge and therefore cannot produce 'appropriate' clinical explanations for their decisions. Charniac (165), however, favoured, justification based upon associations and not first principles as a "Bayesian basis of common sense".

Where belief networks are employed, the model can be used to produce explanations that describe the pathway of cause and effect that has lead to a system's conclusion (136,138).

Van der Lei offered a 'critiquing' model as an alternative to the 'Greek Oracle' when describing HYPERCRITIC, a support system that can audit a general practitioner's management of hypertension. In the system, explanatory comments are available at each stage of the decision making process (166). The system compares coded information entered into the patient record of the general practice computing system with expert guidelines. On many occasions the system was not able to provide advice because records held insufficient and inadequately coded information about the general practitioner's decision making. Problems also arose where experts disagreement concerning the content of guidelines.

Now we are moving into the realms of evaluation where the users and patients have an opportunity to decide whether they would like to have a decision support system (167).

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Chapter 2

Review of Decision Support Systems: Evaluation

1. The Evaluation and Implementation of Diagnostic Advice Systems

The previous section described the evolution of diagnostic advice systems from a system designer's point of view. In the following discussion of evaluation and implementation, the emphasis switches dramatically from the theoretical aspects of knowledge acquisition, knowledge representation and inference model selection to the practical needs of users and patients who might be exposed to such systems.

2. What to Evaluate?

In spite of the wealth of published information that is available concerning diagnostic advice systems, there appears to have been little demand from the medical profession for their implementation.

In 1973, Rosati (168) asserted that,

"physicians will first welcome computer decision aids when they become aware that colleagues who are using the machine have a clear advantage in practice."

Perhaps this will then encourage further implementation (32).

According to Kunz (34), there are two main issues that might effect the implementation of clinical advice advice systems into medical practice;

- The design of effective systems that will help physicians reach better and more reliable decisions.

- Methods of encouraging the use of these systems.

In order to satisfy the first of these requirements there needs to be evaluation of system performance. Lundesgaarde (169) found that only 10% of reported medical knowledge-based systems described in the literature had undergone any form of laboratory testing. Few seem to have made the transition from research tool to viable product.

Wyatt and Spiegelhalter (167) have described an outline protocol for advice system evaluation that is analogous to that that conducted on new drugs. They recommend the following steps;

- Check that there is a clinical problem that can be addressed through provision of the proposed diagnostic support system.

- Perform laboratory tests to assess system safety and the potential benefits of implementation. Included are tests of usability, quality and accuracy of the output and its relationship to expert opinion, investigation of knowledge sources and knowledge representation, power and robustness of the inference mechanism and the reactions of users and patients.

- Perform field trials of the system to assess the effects of intervention upon the conduct and outcomes in clinical care. Included are an explanation the effects of implementation, assessment of system acceptability, any changes in outcome measures and quality of care, and a cost/ benefit analysis. Field trials should be repeated at different centres to eliminate any local bias to the results.

- Conduct post-marketing surveillance of product and publish the results of both successful and unsuccessful trials.

Wyatt applied the format of the protocol when describing the evaluation of the ACORN (Admit Coronary Care Unit or Not) system which is designed to provide advice on the triage of patients suffering with ACP (25). ACORN is an expert system that uses a rule base of some 200 rules and hypothetico-deductive approach to reasoning.

Nykänen (170) highlighted the need to test the behaviour of user/ advice system combination and verify that the software performs the tasks that it is designed to do according to a specification. He went on to suggest that an iterative development and test cycle was more realistic than a single formal trial and that systems should have to undergo regular requalification throughout their lives.

3. The Requirement for Evaluation of Advice Systems

The designers of the Oxford System of Medicine, in disagreeing with the philosophy of the protocol, pointed out that they had adopted an approach to evaluation based upon engineering principles where modelling is used to predict the likelihood success or failure. In order to achieve this they had concentrated their efforts on the maintenance of a theoretically sound inference procedure (146) which had been independently compared with other methods (171).

Other views have been expressed upon what constitutes successful performance in evaluation. Reggia (162), for example, suggested that any system providing even weak justification for its prediction could claim to offer some advantage.

Feldman and Barnett (172) have suggested that formal evaluation of decision support systems might well be a waste of valuable resources, but concede that there is a need to supply users with some feedback on system performance.

4. Problems of Definition

Computer based diagnostic advice systems require the input of medical information before they make any inferences concerning a patient. Many medical terms are inconsistently defined in the literature. Examples of this can be seen even in routine measurements such as blood pressure, where several acceptable ways of determining the diastolic can be found.

Inadequate definition can lead to inconsistent reporting and recording of similar events. For example, Knill-Jones found that when he asked 40 gastroenterologists to define 'flatulence' he received a mixed response (79);

19 believed that the term related to passing wind upwards,
7 to passing wind per rectum and
11 thought the wind could go either way.
3 presumably did not know.

A way of minimising such errors is to define the terms that will be used in an advice system (173,174).

The Clinical Problem

Another important definition is that of the clinical problem that the system is alleviate [1.2.] [1.3.] [1.4.]. For example, Wyatt studied the management of patients who attending a casualty department with acute chest pain in order to determine whether a case could be made for the provision of a decision support system (25). He found that 12% of patients with cardiac disease, who had a high risk of serious complications were being sent home, and that 5% of patients who did not have cardiac disease were being admitted to the coronary care unit. In addition patients were waiting a median of 32 minutes before seeing a doctor and most spent over a hour in the Accident and Emergency department, before being sent home or admitted.

5. Evaluating the Performance of a Diagnostic Advice System

The bounds of the clinical domain in which a system is designed to operate need to be defined to allow testing to be performed using appropriate clinical material according to appropriate clinical standards.

For example, In describing GLADYS, a decision support system for dyspepsia (79), Knill-Jones defines dyspepsia as being;

"Episodic, persistent, or recurrent abdominal pain or discomfort or any other symptoms referrable to the alimentary tract except for rectal bleeding and jaundice as the main symptoms"

None of the large domain commercially available medical decision support systems such as QMR, ILIAD and DXplain appear to have strict domain definition (129,139). There is variable coverage of the medical specialties in both QMR and ILIAD (130). QMR, for example, only contains information about a handful of dermatological diseases, whereas ILIAD contains no skin disease records at all (personal evaluation of ILIAD and QMR, Nov 1992).

6. Laboratory Tests of Case Driven Diagnostic Advice Systems

A critical issue in the evaluation of diagnostic advice systems is the assignment of standards against which their performance can be measured.

In case driven advice systems the standard is commonly the 'true' diagnosis. Provided that we are confidently able to identify that a patient has a particular disease, then we can count the number of occasions that a system has correctly predicted its presence.

Patients who suffer dyspepsia might have one or more of a number of diseases. Knill-Jones was confidently able to identify the presence of a duodenal ulcer in 25% of 1200 dyspepsia cases studied, but in 15% of the total he suspected that the symptoms were caused by irritable bowel disease. Within this group, however, he was only able to confirm the diagnosis on 50% of occasions (79).

For the purpose of evaluation, the identification of disease often relies upon the use of generally accepted definite 'gold standard tests'. For example, in patients who have suffered acute abdominal pain, the pathological finding of an inflamed appendix at operation has been taken to indicate the presence of the appendicitis (5).In acute chest pain the

WHO definition of acute MI, which uses various combinations of history ECG and enzyme findings, has often been acceptable when definitive tests, such as CK-MB estimation have not been available (25,149).

There is not always agreement concerning the definition of a diagnosis or end point. Knill-Jones found that 87% of his patients with dyspepsia had suffered bouts of abdominal pain, but did not have acute surgical disease (79). He was concerned that, by de Dombal's definition, all these patients would be considered to have NSAP, whilst they were actually suffering from a number of diseases that should be treated in different ways.

Card (175) suggested that disease complexes such as irritable bowel syndrome might be defined through the use boolean sets as a representation of the combined opinion of several experts. Wyatt and Spiegelhalter might define this as a 'silver standard' (167). In his description of a Bayesian diagnostic advice system for rheumatological disorders (71), Bernlot Moens related that he had found a paucity of definitive tests for the diseases being considered. He adopted a policy of using experts to estimate the likelihood that each patient had suffered from each of 15 diseases. His 'gold standard' was an independent expert review of each case 6 to 12 months following original consultation.

Accuracy of Diagnostic Prediction

A commonly applied 'laboratory' test of an advice system's function is an estimate of its accuracy of disease prediction. A simple measure is the proportion of cases where the advice system correctly identifies the presence of a disease or diseases in patients whose disease status is known. Bayesian systems produce a relative likelihood result, where the outcome with the highest posterior probability is normally taken as the prediction (5,72,74). If more than one disease can be predicted then more complex comparisons of ranking may be appropriate (139), for example;

Ohmann (69), estimated the 'goodness of fit' of the predicted to actual diagnosis by use of the quadratic (or Brier) score;

$$\frac{1}{N} \;.\; \Sigma \left[(1-Pi_{d(i)})^{2} \;+\; \Sigma_{i<>d(i)} \; P_{ij}^{2} \right]$$

Where N=number of patients, Pij= posterior probability for Dj,
d(i)= index of the actual disease of the patient i
(after Titterington) (176)

Where systems have been derived from case record analysis, three basic methods have been used to create and evaluate the system. The first and least reliable has been to use the whole set of cases both in creation and testing. This produces an over-optimistic estimation of accuracy (74, 167,177).

The second method is to split the cases at some point in their sequence and use one portion as a teaching set and the remainder as a test set. The method is straight forward and normally gives a reasonable estimate of future accuracy (163,178).

Accuracy tests are subject to error of measurement so that in general larger test groups will produce more reliable results, providing that the composition of the test group is appropriate for proposed system use (74). A problem with splitting the collected cases into two groups is that valuable information about cases in the test group cannot be used by the predictive algorithm (74) and there is pessimistic bias with relation to a final system which may be created using all the available case information.

A method of overcoming this is by the computer intensive 'jackknife' or 'one out' principle (48), where the whole case set bar one is used to predict outcome in the remaining case. This procedure is repeated until a prediction has been made for each case. The method has been found to produce a slightly better estimate of future system performance than the training/ test group method (74) and is particularly valuable where the total number of cases collected is small.

There has not always been agreement on what constitutes an appropriate test group. Wyatt and Spiegelhalter recommend that the population studied should represent the one that is to be assisted (167).

A common method of assessing whether a diagnostic advice system might be of value to potential users has been to compare the diagnostic accuracy of the system with unaided user accuracy (4,71). If the system's accuracy is found to be greater then that of unaided users then this supports the hypothesis that the user might increase his chance of achieving a correct diagnosis by referring to the system's prediction. For example, in the evaluation of a diagnostic algorithm for heart disease in neonates (179) a comparison was made between diagnostic accuracy of referring paediatricians (=48%), specialists (=64%), and the algorithm (=78%) working on the same set of patient information. A conclusion reached was that use of the algorithm might reduce morbidity and mortality in neonates.

A finding that the diagnostic accuracy of an advice system exceeds that of potential users is important evidence in favour of its use. However, the statistic may mask vital evidence concerning the quality of the system's output. For example, if a acute chest pain advice system was found to have a higher diagnostic accuracy than clinicians but, it detected fewer cases of MI, then implementation of the system might well be called into question. In situations where it is important to detect a particular disease measures of sensitivity and specificity are often given.

Further examples of poor quality output might include, the misclassification of surgical diseases as non-surgical (5) or malignant tumours as being benign (180).

Laboratory Tests of Case Driven Diagnostic Advice Systems for Acute Chest Pain

A number of case driven diagnostic advice systems have been developed to assist doctors in the identification of diseases causing acute chest pain. Many have undergone laboratory testing.

de Dombal has developed a Bayesian system that can be used to assist doctors with diagnosis in patients presenting with ACP. The approach used was similar to that used in development of the Leeds abdominal pain advice system. Details of history ECG and SGOT levels were used to make predictions of diagnosis from five categories including myocardial infarction, angina, non-specific chest pain, pneumonia and pneumothorax.

Laboratory tests of the system were performed on a total of 973 prospectively collected hospital and general practice records. When the system was assessed for its ability to detect cases of MI, de Dombal found the sensitivity to be

94.6% and the specificity 81.6%.

Goldman (90), tested the Leeds chest pain model on 900 prospectively collected casualty and in-patient records obtained from patients younger than 60 years old with no history of MI. He found the sensitivity of the system in detecting MI to be only 21% and the specificity 90%. He found that his own clinical algorithm, derived by recursive partitioning (Goldman i), performed considerably better on the same test set achieving a sensitivity of 97% and a specificity of 80%. In turn the Goldman method was tested by Poretsky (29) who found that the system's performance compared unfavourably with that of unaided physicians. Goldman subsequently revised his clinical algorithm (90) (Goldman ii) and tested it on a further 4770 casualty department patient records finding the sensitivity in detection of MI to be 88% and the specificity 74%.

Claims of outstanding accuracy in the diagnosis of IHD have been made for two other systems, by Pozen (54) and Joswig (53) that use a linear regression model in prediction. Joswig used his training set of 173 cases to test the performance of the predictive model. Each prediction was compared with the cardiologist's opinion prior to angiography. He found that the system had an overall diagnostic accuracy of 86% compared to that of cardiologists who attained 69%.

Baxt compared the predictive accuracy of a neural network, that had been trained to identify the features of MI, with that of clinicians on 331 prospectively collected consecutive cases of acute anterior chest pain. He found the sensitivity of the physicians in identifying MI (=77.7%) was considerably less than that of the network (=97.2%) and concluded that the performance of his system exceeded that of any other chest pain decision support system and might well prove to be a valuable aid to doctors (149).

In a laboratory test of the ability of the expert chest pain advice system ACORN, to identify which of 174 patients had high risk cardiac disease, Wyatt found the system appeared to be have a greater sensitivity (=75%) than casualty doctors (=62.5%) (25,30).

7. Laboratory Tests of Expert Systems

It has often been assumed that as expert clinicians provide diagnostic and other support to non-specialists, the key to evaluation of expert systems is to show that they behave like experts (181). The measure of similar behaviour can be similar diagnostic performance or similar processes of reasoning. A system that is quoted as being 100% accurate under these circumstances might still be unable to 'correctly' identify a proportion of the cases within its domain.

A variety of criteria have been used to define the standards of expertise against which the performance of expert systems have been measured. Different views have also been expressed concerning the selection of appropriate test cases, particularly where systems have been designed to provide assistance in difficult clinical case. Here, Wyatt and Spiegelhalter suggest a random mix of difficult and routine cases (167).

Relatively few assessments have been quantitative (72). For example PUFF was considered to give good performance because on 90% of occasions, its advice did not need to be altered by a supervising expert (114) (how many of these records were normal?).

An expert system for thoracic pain diagnosis has been developed by Puppe (182). The model uses approximately 1000 diagnostic rules in order to differentiate between 18 possible disease categories that include, amongst others, the common presentations of IHD. The system operates a system of hypothesis generation and evaluation as a model of human decision making. The author conducted an evaluation of system performance based upon simulated patient records and reported that it,
"proved to be quite competent in its domain", although no specific results were given. It was reported, however, that the system could justify its conclusions by listing rules that explained how they were reached.

DXplain is a dial-in diagnostic and management advice system which has a knowledge base that contains information about 2100 diseases and 4500 medical terms (172). In an evaluation of the system a total of 65 test cases were drawn from three sources; users, expert physicians, and reports in journals in order to achieve a balance that reflected the variety of problems encountered in clinical practice. Cases were only accepted, however, if the diagnosis was represented in the knowledge base. In the performance test, the expert panel could only use clinical information that could be coded for system use. For each case, the differential produced by DXplain was compared with the majority view of a panel of 4 experts and the system itself. A conclusion reached was that DXplain was behaving like an expert (172).

A recent test of the performance of QMR, used a series of case summaries created by experts according to guidelines (139). The experts submitting cases provided a 'usefulness score' for each item of patient information that indicated its relevance to diagnosis. Assessment of the reasoning ability of the system was made by weighing up the positive and negative scores for collected evidence. Cases were discarded, however, if the true diagnosis did not appear in the knowledge base.

In a test of ELIAS, a system for auditing general practitioners' decision making about patients with hypertension, the comments about each patient generated by the computer were compared with the views of a panel of experts, with the system being considered as having produced the correct result if its output agreed with the opinions of 6 out of 8 members of the panel (166). One conclusion was that for at least some of the time experts appeared to be making arbitrary judgements about what the correct advice or course of action should be.

In an evaluation of a neural network trained to identify malignant breast calcificaton, test cases were those considered difficult by experts. The accuracy of experts was assumed to be 50% and the network to be of potential value because it was able to correctly identify 72% of cases where malignancy was present (183).

Pople (184), expressed the opinion that CADUCEUS (nee INTERNIST) does a good job. It had been tested on 43 clinical cases of which it had successfully identified the diagnosis in 17 cases. In comparison experts had been able, on average, to get 23 of the 43 correct.

Yu (185), compared the performance of clinicians with MYCIN and found that 65% of the output of MYCIN was acceptable as compared to only 55.5% of that of clinicians. Teach (33), appeared to be unimpressed with the validation of the rule set for MYCIN which he reported had been tested on 15 patients.

The advice produced by MYCIN has since been compared with that of experts in a blinded laboratory evaluation by a second set of infectious disease experts. However, the test was only concerned with therapy selection in a single disease, meningitis (186).

8. Requirements of Users

"The art of medicine consists of amusing the patient (? doctor) while nature cures the disease"

(Voltaire, 1694-1778)

Decision support systems should meet a medical need with an appropriate solution (187).

de Dombal has suggested (188), for example, that teaching packages might be just as good, and less threatening, as on line decision support in improving the diagnostic performance of junior surgeons.

Systems should be supplied to those who need them. The majority of advice systems have originated in and have been designed for use in hospital departments. Wigerz (39) discussed the importance of providing decision support for primary care as this is where most generalists work and decision support might be most effective.

Attention to the provision of a suitable terminal and interface (83) will promote compliance. Systems are also more likely to be used if their use causes minimal disruption to normal routine (32), particularly to consultation times (189), and if they are integrated with existing systems (170,190).

A decision support system should be able to provide its advice at the correct point in a consultation. In a field test of ACORN, it was found that in (25) 25% of test cases, the system was actually used too late to effect decision making.

As well as being accurate, advice given by a system should be appropriate. The user should not be burdened with conducting numerous additional and unnecessary tests (191,192).

The potentially deleterious effects of system use during consultations upon the doctor patient relationship must also be considered (170). This problem was side-stepped in trials of the Leeds AP system by using a clinical assistant to enter onto the computer, patient data that had been collected by house officers. The clinical assistant then returned output advice to the requesting doctor (5).

Machine expertise in medical diagnosis is a sensitive subject amongst medical practitioners, (170). Resistance to implementation is likely to be greatest if it is suggested that doctors should be subservient to 'expert' computers. Doctors have no desire to be replaced by computers nor do patients want this (72). In any case, it seems likely that doctors will remain legally responsible for care given to their patients, whether on not they receive advice from decision support (1) systems.

Clinicians might be interested in making use of systems that provide feedback about doctor performance including accuracy of diagnosis (5). They might also take an active interest if they are invited to become involved in the development cycle (193).

Wyatt and Spiegelhalter have suggested that the responses of users and patients to system implementation during evaluation could be sampled by the use of questionnaires (167). The designers of the Oxford System of Medicine (OSM) seem to favour a pre-emptive approach. As part of the LEMMA project, the OSM was demonstrated to 200 general practitioners who were then asked to comment on the suitability for implementation of various modules within the system. The conclusion was made that general practitioners approved of the proposed development plans (194).

9. Field Trials of Medical Decision Support Systems

Wyatt and Spiegelhalter concluded that field trials of decision support systems are necessary (167), not least to indicate whether systems will actually be used and to obtain feedback concerning how their use will affect the practice of medicine (195). They suggested that a double blind randomised controlled trial methodology should be adopted for the conduct of field trials, where the control and intervention groups are matched apart from the availability of advice. Confounding factors such as the Hawthorne and 'checklist'(179) effects should also be taken into account.

The 'checklist' effect is an improvement in diagnostic accuracy brought about by using a structured list of relevant questions to ask about a condition. It seemed to account for some of the improvement in performance found in house surgeons taking part in trials of the Leeds AAP system (5,8). In another example, an evaluation of a diagnostic algorithm for detecting heart disease in neonates, the diagnostic accuracy of both paediatricians and specialists improved by about 10% when they used structured questionnaires for collecting information about new patients (179).

10. Field Trials of Case Driven Diagnostic Advice Systems

a) The Leeds Acute Abdominal Pain Advice System

In the late 1970s de Dombal and his co-workers proposed a clinical application for the Leeds acute abdominal pain system. He suggested that junior doctors' diagnostic accuracy might be improved through 'real time' use of the program and that it might be possible to measure the effects of these changes upon the provision of health care. Increased diagnostic accuracy might, for example, lead to a reduction in the number of patients admitted with suspected appendicitis, who were eventually found not to have the disease (4,7).

A tri-phasic method of clinical testing was developed and tested (7);

Trial Methodology

(i) The initial pre-intervention phase involves collection of information, about the performance of surgical teams, that can be used as a baseline for further comparison. Performance indicators that have been used include admission and discharge rates, diagnostic accuracy by grade of surgeon and details of operations performed. Monitoring of performance continues throughout the phases of the trial.

(ii) In the second phase, house officers are encouraged to use abdominal pain data collection sheets during clerking.

(iii) In the final phase, the acute abdominal pain program is made available and can offer diagnostic advice based upon collected data.

A pilot study produced encouraging results (7) and lead to government funding being made available for a multi-centre trial of an upgraded system that included a more extensive database. The trial methodology also evolved and included additional cross-over phases, where intervention measures were removed in order to investigate any training effects of system implementation. Leeds workers coordinated trials that were carried out in eight UK hospitals and which involved some 250 doctors and 16737 patients. The results were published in 1986 (5,8,9).

Results of Muli-Centre Field Trials

The main findings of the multi-centre trials can be summarised as follows;

(i) As in previous trials, there were found to be differences in the overall baseline diagnostic accuracy rates between house officers (=46%), senior house officers (=58%) and registrars (=69%) managing patients admitted with acute abdominal pain.
The overall accuracy of the computer when tested on all cases for which forms were completed was found to be 68%.

(ii) The combined interventions of data collection sheet and computer program use conferred, on average, a 15% increase in doctor diagnostic performance irrespective of grade. In breaking this down, it was estimated that 10% could be attributed to the use of forms and 5% to the additional assistance provided by the computer and feedback of personal performance figures.

(iii) When the baseline and intervention phases were compared with regard to performance indicators it was discovered that implementation of components of the system had resulted in benefits for both the patients and hospitals concerned. With regard to patients who were admitted with suspected appendicitis, there were reductions in the total number of laparotomies performed, the proportion of laparotomies that yielded no abnormal findings (negative laparotomy), the appendix perforation rate and the average length of stay in hospital.

(iv) A survey the doctors' responses to system implementation was conducted. This majority appeared to be in favour of further use of the system in practice.

It was concluded that improvements in diagnostic performance of surgical teams using the advice system had been responsible for the measured improvements in surgical care.

The credibility of this conclusion was enhanced by evidence that the improvements in doctor performance and surgical practice could be reversed by removal of the system (5), and that repetition of the effect had been achieved in different centres (8,9), although analysis of the results had been centralised.

b) Field Trials of Acute Chest Pain Advice Systems

Pozen (54) has produced a linear regression model for the prediction of diagnosis in patients suffering with ACP. It has been extensively tested on 2320 new patients in a six centre hospital trial in the USA where it was implemented on programmable calculators. It is reported that its use resulted in a reduction in the proportion of patients with non-cardiac chest pain admitted to the CCU from 44% to 33% (204,216).

11. Field Trials of Expert Systems

Expert systems seem often to have been evaluated to the designers satisfaction in controlled environments using simulated clinical material (72,139). In many cases medical expert systems have been unable to function in the 'real time' clinical environment and have been relegated for use as research tools rather than decision support aids (39,83). A notable exception was HELP which found widespread clinical use (175) as long ago as 1983. Pryor described the evaluation (123);

'the pharmacy module has resulted in an overwhelming positive response form the medical staff who make few medication-prescribing errors. The computer allows them to conduct their practice with the assurance that the problems will not be a major part of their risk in caring for the patient'

QMR, DXplain, ILIAD and RECONSIDER are large domain medical expert systems systems. Both QMR and ILIAD are available commercially and licences for their use have been sold to a total of more than 2500 sites. (196,197,198,199,200).

In an evaluation of ILIAD using 50 consecutive grand-rounds cases, Heckerling found that only 28 of the diseases suffered by the patients appeared in the knowledge base (130). In order to predict the effects of ILIAD's use in practice he went on to compare the differential diagnostic lists produced by two internists before and after they had consulted the advice system. He found that overall the mean ranked position of the correct diagnosis improved after

ILIAD's advice but that internists' diagnostic accuracy did not change. He commented that the system did not appear to recognise when it was dealing with a disease that was not in its knowledge base.

The large domain medical systems such as QMR, DXplain, ILIAD and RECONSIDER appear to have incomplete knowledge bases and to have been subjected to extremely limited field trial evaluation, but at least three are commercially available and in widespread routine use in the USA. The OSM is another large domain Expert system which the designers hope will evolve into a "European System of Medicine" (201). They apparently disagree with the principle of conducting blinded controlled field trials before implementation (202).

Wyatt has conducted a blinded and controlled field trial of the expert system ACORN in a casualty department to determine the effects of implementation of the system upon the staff's clinical management of patients attending with acute chest pain (25). Patients were randomly allocated to either an ACORN use or control group following data collection by a nurse. Of the 153 cases admitted to the study 14 were excluded because a gold standard diagnosis could not be determined.

The casualty officers' diagnostic accuracy was compared between control (=92%) and intervention (=90%) groups and the conclusion drawn that implementation of the system had not produced a beneficial effect. Little difference was found in false positive and false negative rates for the prediction of high risk patients between the groups and such patients also appeared to be waiting longer before admission. It is of interest that the accuracy of diagnosis of patients in the control group was higher than that found in a baseline study of casualty officer performance (25).

Despite the adverse results of the carefully conducted and controlled field trial of ACORN, the system was modified and continued to be used in the casualty department after the trial had finished (25)

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Plan and Justification of Experimental Work

The Study of Case Driven Diagnostic Advice Systems
Scope of the Thesis

The three main sources of clinical information that have been available to the designers of diagnostic advice systems have been experts, publications and patient records. Unfortunately the information obtained from experts has often been found to be vague, incomplete, unreliable and inconsistent (210,211) [1.9.b].

Textbooks and journals contain variable amounts quantitative and qualitative information. These can range from numeric derivatives of population study to incomplete and general (210) subjective impressions couched in such vague terms as "seldom" or "frequent" (43). "Typical case" descriptions are also commonly found in standard texts.

Obtaining information from patient records is time consuming and the results can be subject to errors introduced through observer variation, inaccurate transcription, missing information and unclear definition of terms and end points. Retrospectively collected databases are less likely to be accurate that those that have been prospectively collected due to missing information and subjective variation in definition between examiners (33).

The various types of inference model available are suitable for manipulating particular sorts of data. The designers of case driven advice systems have tended to use iterative and statistical methods, whereas those relying upon expert opinion have on the whole adopted expert system techniques. There has been cross-over where statistical inference has been used to fill gaps in, and tune, represented expert knowledge and beliefs (128) [1.9.d]. This, of course, represents a partial reversion to prospective data collection (210).

"The only source of valid data for computer-assisted decision making systems is a carefully recorded, adequate, prospectively collected survey group" (de Dombal, 1983)

The notion that decision support systems need to model human knowledge and reasoning leads to problems as human expertise is expensive, difficult to define and not necessarily available (170) [1.8]. Modelling of the human decision making process does not appear to be an essential criterion for successful matching of diagnostic accuracy rates by advice systems as many simple mathematical predictive systems have demonstrated (72). Simple advice systems might alter attained clinician accuracy by eliminating subjective bias or by providing hard statistical back up, or cues, for decision making (72).

It appears that attitudes to the methodological differences between proposed diagnostic advice systems, have to some extent delayed the production and introduction of potentially efficient clinical aids that are problem rather than method orientated [1.6.]. There has been a tendency, over the years, to produce increasingly complex representation and inference models in order to cope with the theoretical disadvantages of those that have preceded them [1.9.]. There has been a progressive change in emphasis from case driven to 'expert rule' to 'expert belief' driven systems. Yet, there is little evidence from evaluation to suggest that this evolutionary path is producing clinical systems that are any more able to favourably influence clinical care than their ancestors (34) [1.12] [2.11]. The important question to be answered is perhaps not as asserted by Carroll (72); "what is the best model of clinical decision making?", but rather "how can these models best be used to support and modify outcome (99)?".

In the light of presented evidence, my conclusions from this line of reasoning were that in the development of a new diagnostic advice system, I should;

- use carefully and prospectively collected patient information as a primary source of knowledge about disease in order to minimise selective and interpretative bias that might otherwise be introduced through sole reliance upon expert opinion

- chose a medical domain where a requirement for decision support could be identified, there was a high through-put of cases and where disease end points could easily be determined

- concentrate, initially on the use of simple inference models for diagnostic prediction

- be prepared to select, adapt or reject particular inference models in the light experimental evidence concerning their suitability for the chosen decision task.

The choice of dermatology as a medical domain for new system construction was influenced by the following factors;

- dermatology diagnosis appears to cause problems for non-specialists

- the manifestations of dermatological disease are visible and therefore available for description by an observer collecting information.

- a standard nomenclature exists for describing skin lesions

- diagnosis can be confirmed by sampling exposed tissue

- dermatological disease is sufficiently prevalent in the community to allow rapid accumulation of case data and warrant construction of an advice system.

For these reasons I decided to design, construct and evaluate a case driven advice system that could assist general practitioners with the task of dermatology diagnosis. My hypothesis was that such a device might improve the accuracy with which diagnosis was made and lead to improvements in patient health care.

Experimental Work Performed; Requirement, Nature and Extent

1) Comparison of Inference Models for Acute Chest Pain Diagnosis

There is evidence that the choice of inference model has less influence upon predictive ability of case driven diagnostic advice systems than other factors such as the type quality and completeness of data (80,89,90,91). This might account for the finding that a number of case driven and other models have been recommended for use in assisting doctors with the identification of patients suffering ischaemic heart disease. However, the implemented systems use different sets of variables (53,54,89,90,182,195,214, 216).

The purpose of the investigation has been to carry out an independent comparison of several established acute chest pain diagnostic advice systems in order gather information concerning the relative performance and applicability of different inference models used in the same clinical setting.

2) Hospital Trial of The Leeds Acute Abdominal Pain Diagnostic Advice System

The Leeds acute abdominal pain diagnostic advice system has probably been more thoroughly evaluated in laboratory and field testing than any other medical decision support system and as such has been selected as a benchmark for comparative study and further investigation [1.10.a]. There are several outstanding issues;

- the system has been shown to confer advantage upon its users and patients yet it has not been accepted into routine clinical use

- surprisingly, there has not been true independent evaluation of the system as all field trial results have, to date, been processed by the design unit

- in field trials, the diagnostic performance of the computer appeared to fall below that of doctors that had been assisted by it. The mechanism by which doctors could have been assisted by the computer under these circumstances does not appear to have been investigated.

- little use has been made of decision justification routines in implementation other than the production of standard textual summaries. Provision of more extensive justification might improve user acceptance of the system.

The field trial has been carried out in order to shed light on these issues and in order to gain sufficient practical insight to the applied trial methodology to allow comparison with that recommended by Wyatt and Spiegelhalter (5,7,167).

3) Comparisons of the Performance of The Leeds Acute Abdominal Pain Diagnostic Advice System with Paramedics, Non-Medical Staff and Referring General Practitioners

There has been little investigation of the possibility of using the Leeds advice system in primary care where non-specialists might welcome diagnostic advice when making decisions concerning patients suffering with AAP.

Laboratory trials have been carried out in order to assess the potential value this advice system to general practitioners, paramedics and other personnel charged with providing health care. The trials have also allowed assessment the implications of moving a support system designed for use in secondary care to a primary care setting.

4) The Design and Construction of DERMIS: A Primary Care Advice Dermatology Diagnostic Advice System

The Leeds acute abdominal pain advice system uses a simple Bayesian diagnostic inference model. Variants of the technique continue to be popular amongst diagnostic advice system designers. The method is easily understood and appears to be reasonably robust and transportable between domains (67) and when it has been compared directly with other more sophisticated models it often appears to perform just as well [1.7][1.12](72)(143). It has been adopted for use in initial development of an advice system for dermatology.

The following design and construction work has been carried out;

- investigation of the diagnostic accuracy and referral patterns of general practitioners managing patients with skin disease

- prospective collection of clinical records to construct a database

- implementation of a Bayesian inference model on computer

- practical comparison of several methods of data entry

5) Investigation of Measures that Can be Taken to Improve the Performance of Diagnostic Advice Systems that Use a Simple Bayesian Model

The theoretical disadvantages of simple Bayesian inference are well known [1.7.d] [1.7.e] [1.14]. For this reason, known weaknesses and application problems of the model have been investigated during diagnostic performance tests in order to determine they are actually of clinical and practical importance within the selected domains. Particular investigation has been carried out in the following areas;

- the assumption of independence of variables
- the assumption of a mutually exclusive and exhaustive set of diseases
- difficulty in estimation of lower frequncy bounds
- the representation and incorporation of expert beliefs
- justification of results

6) Laboratory Tests of the Performance of the DERMIS System

The diagnostic accuracy has been tested and then re-evaluated in the light of changes suggested by investigative work performed at [2.5] above.
Changes in configuration that have been tested include;

- a reduced data set
- grouping of diseases by pathological process and equivalent treatment
- comparison of 3 lower frequency bound estimators
- inclusion of expert beliefs concerning the occurrence of features in diseases.

7) Trials of the DERMIS System in Practice

Semi-field testing of the DERMIS system has been carried out in order to investigate the potential problems and implications of implementing the system in primary care. The following issues have been addressed;

- the choice of user-interface
- expert review of system performance
- the effects of observer variation in data collection upon system performance
- the effects of system use upon diagnostic and management decision making.

The methods adopted in investigation, results obtained and further discussion of these issues are presented in the remaining sections of this thesis.

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Chapter 3

Experimental Work: Methods

Methods Applied in Experimental Work

The plan of work for this thesis has been described. The methods employed to carry out the work are presented in this section in the same order as the tasks appear in the plan of experimental work [page 90]. Where the same method has been used on more than one occasion, it is described in detail only at its first usage.

1) Comparison of Inference Models for Acute Chest Pain Diagnosis

a) Description of the Inference Models to be Tested

(i) Bayesian Advice System : de Dombal

de Dombal (63) has described a system that uses history, examination, ECG and optionally cardiac enzyme (SGOT) results in order to predict patient diagnosis and prognosis. In the version available for test (63), four diagnostic categories are considered, including myocardial infarction, angina, non-specific chest pain and chest infection. The user completes a standard questionnaire which has 47 multi-faceted questions giving a total of 175 possible options.

An algorithm based on Bayes theorem is used to compare new case details with a database of frequencies compiled from prospectively collected clinical cases in order to produce a relative likelihood output which can be interpreted as a differential diagnosis.

The system is designed to be used for triage of acute chest pain cases which are assigned to one of three risk groups according to relative likelihood score (64,212).

A Simple Bayesian Algorithm for Practical Use

Practical use of Bayes's rule (Thomas Bayes 1702-61) can be made as follows (60,80,137)

A patient may be described by the complex (D,S)
where
D is a set of J diseases which are labelled d1-dJ
S is a set of K features that can be used to describe D
and which are labelled s1-sK

d1-dJ are assumed to be mutually exclusive and exhaustive for the chosen domain

D_{1-J} distinct diseases

symptoms S=(s1-sK) most applications facet s polychotomous
one of a finite list
(assume sensible classes)

patient is (D,S) we observe S and predict d given that the patients set of features = S

Probability that {D=dj'|S} = $\dfrac{P(dj')\ P(S|dj')}{\sum_{j=1}^{J} P(dj)\ P(S|dj)}$

(D,S) complex is enormous so we assume CI (213,3,62)

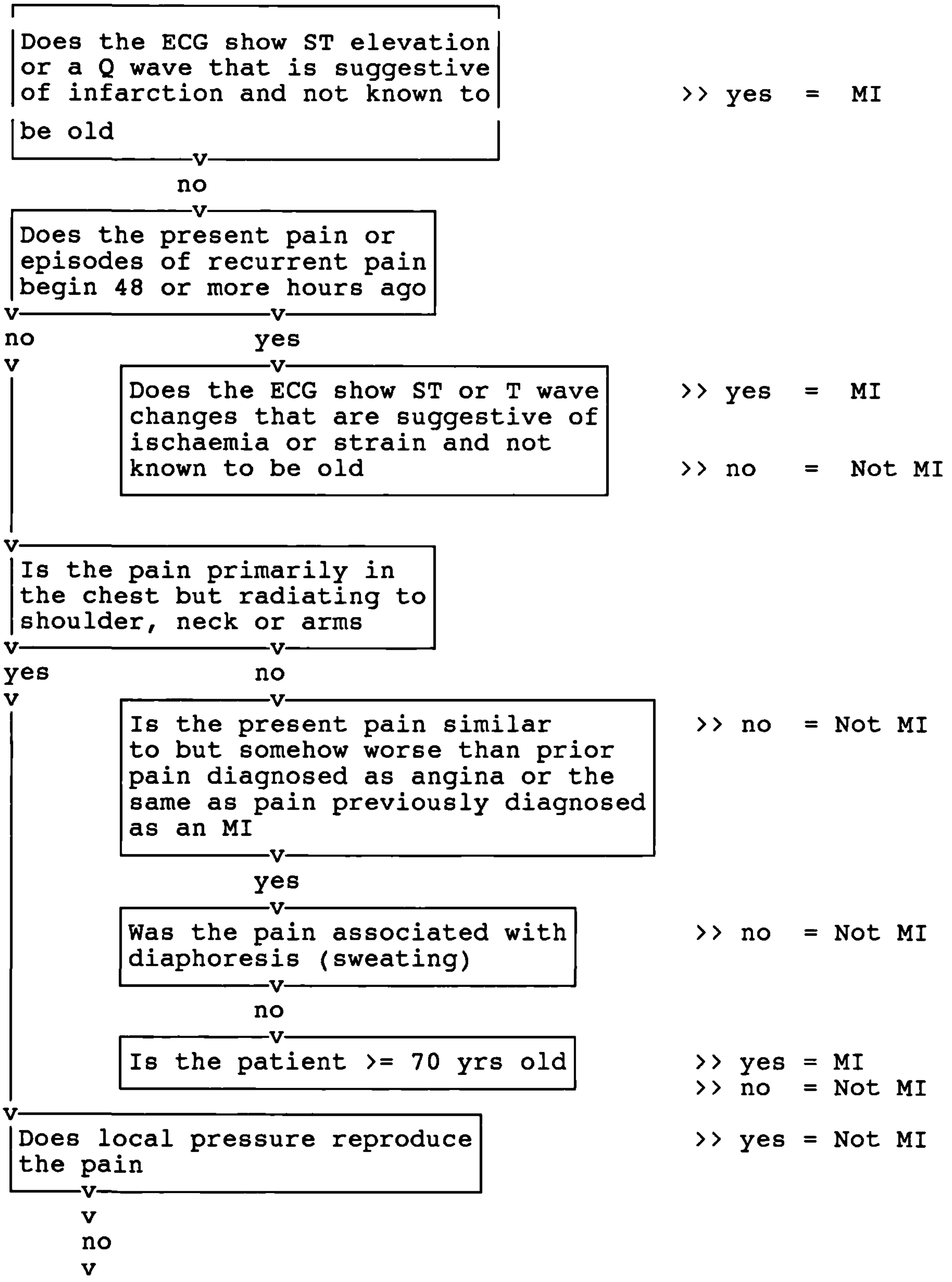
$$P(S|dj)=\prod_{k=1}^{K} P(sk|dj)$$

Where P(sk|dj) represents a frequency estimate produced by an expert or derived from population study.

(ii) Decision Tree Derived by Recursive Partitioning : Goldman

Goldman has used recursive partitioning to develop two decision trees for use by casualty officers managing patients suffering with ACP. The earliest general form of the model is summarised below (89)

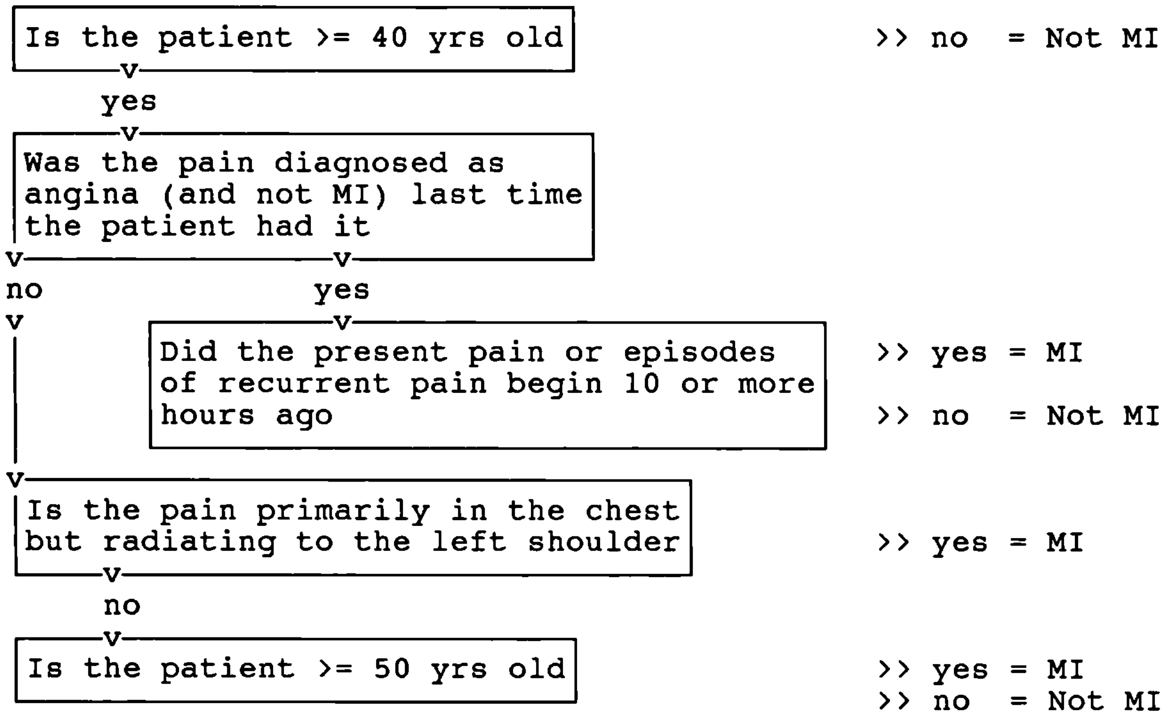
Figure 2. Goldman Recursive Partitioning Tree (i) Model for ACP Diagnosis



```
Does the ECG show ST elevation
or a Q wave that is suggestive
of infarction and not known to             >> yes  =  MI
be old
            v
            no
            v
Does the present pain or
episodes of recurrent pain
begin 48 or more hours ago
v               v
no              yes
v               v
|    Does the ECG show ST or T wave        >> yes  =  MI
|    changes that are suggestive of
|    ischaemia or strain and not
|    known to be old                       >> no   =  Not MI
|
v
Is the pain primarily in
the chest but radiating to
shoulder, neck or arms
v               v
yes             no
v               v
|    Is the present pain similar             >> no   = Not MI
|    to but somehow worse than prior
|    pain diagnosed as angina or the
|    same as pain previously diagnosed
|    as an MI
|               v
|              yes
|               v
|    Was the pain associated with            >> no   = Not MI
|    diaphoresis (sweating)
|               v
|              no
|               v
|    Is the patient >= 70 yrs old            >> yes = MI
|                                            >> no  = Not MI
v
Does local pressure reproduce                >> yes = Not MI
the pain
      v
      v
      no
      v
```

Figure 2 (continued)

Figure 2. (Continued) Goldman Recursive Partitioning Tree (i) Model for ACP Diagnosis



Is the patient >= 40 yrs old

>> no = Not MI

yes

Was the pain diagnosed as angina (and not MI) last time the patient had it

no

yes

Did the present pain or episodes of recurrent pain begin 10 or more hours ago

>> yes = MI

>> no = Not MI

Is the pain primarily in the chest but radiating to the left shoulder

>> yes = MI

no

Is the patient >= 50 yrs old

>> yes = MI

>> no = Not MI

The diagnostic tree has been recommended for use in early diagnosis of patients with ACP (48), who have a clear chest X ray and no immediate history of chest injury. In order to produce a prediction of diagnosis, the questions are answered sequentially until a MI/Not MI conclusion is reached.

The 1988 version of the protocol (90) (Goldman model ii) follows the general format of the earlier model.

(iii) Joswig: Logistic Regression Model

Joswig (53) has devised an advice system based upon the logistic regression analysis of the responses found in a prospective study of ACP patients sent for coronary angiography. Information was collected from 184 patients using a standard history and examination sheet which had 32 multi-faceted questions that gave a total of 157 possible options. In order to fit the regression model, the number of variables was reduced to 13, by evaluating Chi-squares and rejecting those that did not have a significant association with outcome. For practical use (54), a calculator can be programmed to accept the yes/no (1 0) answers to the questions and predict the probability of coronary artery abnormality.

The result of the logistic regression was the following subset of 13 variables and their coefficients (and a constant term) shown in [Table 1]:

Table 1
Joswig Logistic Regression Model: List of Features and Coefficients

Variable	Coefficient
1. Sex of patient	-0.915
2. Age of patient	-0.100
Type of pain	
3. Burning	-0.568
4. Prickling	0.983
5. Radiates arms	-0.519
6. Aggravated movement	1.747
7. Aggravated by sex	-1.142
8. Dyspnoea with pain	0.404
Other	
9. Nausea	0.425
10. Diabetic	-1.432
11. Elevated lipids	-0.458
12. Abnormal ECG	-0.899
13. Pain relieved by rest	-0.495

When examining a new case, the questions are answered in turn and the response assigned a '1' if positive or '0' if negative. The answers are entered into the equation;

$$T = C + \sum_{s=1}^{s=13} (Rs . Coeff(s))$$

in order to obtain

$$P(IHD|S) = \frac{1}{1 + Exp(T)}$$

Where P(IHD|S) is the probability of IHD given the feature (symptom/sign) set S is the set of 13 predictive features (s= 1 to 13). C is the constant term of regression. R is the response variable (R=1 or 0) and Coeff (S) is the set of feature coefficients of regression.

Joswig recommends that a P(IHD|S) value of 0.5 or greater should be taken as indicating the presence of IHD.

(iv) Pozen: Logistic Regression Model

Pozen (54) has also used logistic regression analysis in order to produce an advice system that assesses the likelihood of acute IHD being present. His model, the result of logistic regression analysis, uses the following 7 variables [Table 2]:

Table 2.
Pozen Logistic Regression Model: List of Features and Coefficients

Variable	Coefficient
1. Presence of chest pain	0.9988
2. Chest pain most important symptom	0.7145
3. History of MI	0.4187
4. History of GTN use	0.5091
5. ECG; ST elevation or depression >= 1mm	0.7628
6. ECG; abnormal ST straightening without more than 0.5mm depression	0.8321
7. T wave peaked or inverted >= 1mm	1.1278

A probability value for the presence of acute IHD can be calculated using the response variables, coefficients and the constant term. Wiegert (214), recommends that a probability of 0.4 or above should be taken to indicate an urgent requirement for admission to the coronary care unit.

b) Subjects and Data Collection

Subjects included 108 consecutive ACP patients that were admitted to the CCU at the West Middlesex Hospital, London, over a four month period in 1987.

A data collection proforma was designed which incorporated the items of history and examination required by each of the chest pain advice system models that were to be tested [Table 1] [Table 2] [Figure 2]. In practice, a proforma was completed by a cardiology registrar for each patient admitted to the CCU with acute chest pain and checked by a second experienced clinician. A 12 lead ECG recording was made at the same time and was the source of information for required trace measurements and subjective comment required by some of the predictive models.

Data collection sheets were later checked for inconsistency and errors of omission. Corrections were made from the clinical records where reliable information was available from other sources. Four patients were excluded because their clinical condition prevented data collection.

c) Diagnostic Classification

Myocardial infarction was judged to have occurred in patients with a history of chest pain and development of Q waves, typical enzyme changes or sudden death within 72 hours of admission. S-T and T wave changes were only taken to indicate infarction if there were changes in cardiac enzymes or evidence from cardiac imaging. Persistent residual S-T depression or T wave changes without enzyme changes were classified as sub endocardial infarction.

A diagnosis of angina was made in patients with clinical features suggestive of IHD with transient S-T changes or a positive stress ECG. Non-ischaemic chest pain was diagnosed when angina could be excluded by a definite alternative source of chest pain or an atypical history with a negative stress ECG.

Timed sequential CK-MB analysis was performed for all patients. This information along with clinical, ECG, enzyme,echocardiography, technetium^{m99} multiple uptake gated pool studies, where performed, and ECG analyses were used by the CCU physicians to assign final diagnosis in each case.

d) System Comparisons

Once all the data had been collected and validated, the diagnostic models were tested simultaneously by computer. The performance of each of the described models has been compared with each of the remainder, where appropriate. Comparisons have also been made between the predictive accuracy of the models, casualty officers and ECG reading by cardiologists.

2x2 tables have been constructed and used to estimate sensitivity, specificity and diagnostic accuracy. Mc Nemar's test has been performed in order to estimate the significance of any differences in performance detected.

Comparison of Advice Systems and Advice System Users: 2x2 Tables

From time to time, in the course of the practical work, the ability of various advice systems to predict diagnosis has been measured. Systems have been compared with one another and with potential users. The performance data can be simply displayed using 2x2 tables [Figure 3].

Figure 3.

Assessment of Advice System Performance: 2x2 Table Analysis

'True Outcome'
(Gold Standard)

		Positive	Negative	Total
For System Tested	Positive	a	b	a+b
	Negative	c	d	c+d
		a+c	b+d	n

Where;

a=True Positives (TP), b=False Positives (FP)
c=False Negatives (FN), d=True Negatives (TN)

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Overall Accuracy} = \frac{TP + TN}{n}$$

2x2 Contingency Tables
For comparison of two advice systems working on same cases;

		Advice System B	
		Correct	Incorrect
Advice System A	Correct	a	b
	Incorrect	c	d

McNemar's test for discordant pairs (61) is used with the hypothesis that the two advice systems have the same predictive accuracy.

Where;

$$\chi^2 = \frac{(b - c)^2}{(b + c)}$$

with one degree of freedom

2) Hospital Trial of The Leeds Acute Abdominal Pain Diagnostic Advice System

The conduct of the field trial followed a standard three phase evaluation methodology [1.10.a].

a) Subjects

The Hospital subjects included patients admitted to RNH Haslar, during 1984 and 1985, with acute abdominal pain of less than one weeks duration. The trial was supported by the surgical division and organised into three parts, each planned to be of 6 months duration.

b) Performance Measures

The activity of the surgical department was monitored throughout the trial by collecting the following data set for each patient;

(i) Diagnostic category assigned by the;

- referring doctor
- surgical house officer
- senior house officer / registrar

(ii) Diagnosis after follow up

(iii) Computer advice, diagnostic category achieving highest relative likelihood in a simple Bayes calculation

(iv) Hospital 'performance indicators';

- admission rate
- length of hospital stay
- operation(s) performed and result.

Where possible, the general practitioner's letter, casualty card or Service admission summary was used to establish the referring doctor 'diagnosis'. In many cases, where these papers were not available, the documents were searched for any notes of general practitioner/house officer telephone conversations about diagnosis at referral.

Hospital doctor diagnosis was obtained directly from notes made at the time of admission. Medical officers were advised that they should attempt diagnosis in each admitted case and indicate their choice clearly in the record. They were informed that, for the purposes of the trial, if they gave a differential list, the first disease appearing would be recorded as house officer diagnosis. Where the AAP proforma was used diagnosis was taken from the 'initial diagnosis and plan' box.

Final diagnosis was assumed to be the consultants post-admission summary diagnosis unless follow up over at least one year revealed some other cause. Where possible, included cases were coded according to the computer advice systems' nine main categories. For example, a NSAP diagnosis may have resulted from such diverse problems as dysmenorrhoea, cystitis or mild gastroenteritis. Admission rate and length of stay were calculated from returns and corroborated using the hospital records system. The findings at operation were noted and included such details as, for example, any perforation or lack of inflammation of the appendix. Presumptions of diagnosis made at operation were confirmed only if supported by the results of histological analysis.

c) Baseline Phase

The first part of the trial formed a baseline performance study with which the effects of interventions in the second and third parts could be compared. The staff carried out normal procedure but passed on identification details of patients admitted to the investigator, the present author, who then followed up the patients progress by accessing records after patient discharge.

d) Phase 2: Use of Data Collection Sheets

During the second six months house officers were encouraged to to use the Leeds acute abdomen proforma, and associated definitions (206), when collecting patient history and examination details. These forms were made freely available and when completed were acceptable to the consultants as 'abdominal' clerking, Details of other systematic enquiry were recorded separately. A supply of AAP proforma was made available at all places where clerking might occur and a mechanism was set up to ensure that one copy was returned to

FIGURE 4

Abdominal Pain Chart

NAME

REG NUMBER

MALE/ FEMALE AGE

FORM FILLED BY

PRESENTATION (999, GP, etc)

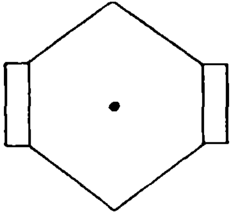
DATE

TIME

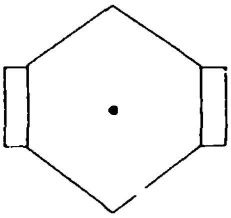
PAIN

SITE

ONSET



PRESENT



RADIATION

AGGRAVATING FACTORS
movement
coughing
respiration
food
other
none

RELIEVING FACTORS
lying still
vomiting
antacids
food
other
none

PROGRESS
better
same
worse

DURATION

TYPE
intermittent
steady
colicky

SEVERITY
moderate
severe

HISTORY

NAUSEA
yes no

VOMITING
yes no

ANOREXIA
yes no

PREV INDIGESTION
yes no

JAUNDICE
yes no

BOWELS
normal
constipation
diarrhoea
blood
mucus

MICTURITION
normal
frequency
dysuria
dark
haematuria

PREV SIMILAR PAIN
yes no

PREV ABDO SURGERY
yes no

DRUGS FOR ABDO PAIN
yes no

♀ LMP

pregnant

Vag. discharge

dizzy faint

EXAMINATION

MOOD
normal
distressed
anxious

SHOCKED
yes no

COLOUR
normal
pale
flushed
jaundiced
cyanosed

TEMP PULSE

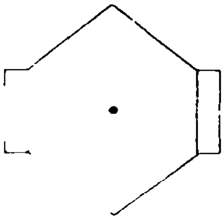
BP

ABDO MOVEMENT
normal
poor nil
peristalsis

SCAR
yes no

DISTENSION
yes no

TENDERNESS



REBOUND
yes no

GUARDING
yes no

RIGIDITY
yes no

MASS
yes no

MURPHY S
+ve ve

BOWEL SOUNDS
normal absent + – +

RECTAL — VAGINAL TENDERNES
left
right
general
mass
none

INITIAL DIAGNOSIS & PLAN

RESULTS
amylase
blood count (WBC)
computer
urine
X-ray
other

DIAG & PLAN AFTER INVEST

(time)

DISCHARGE DIAGNOSIS

History and examination of other systems on separate case notes

the investigator after use. The original proforma became part of the case record.

The trial objectives and methods were explained to participating medical officers and meetings were arranged, where necessary, to review progress and resolve logistic problems.

e) Phase 3: Use of the Computer Advice System

For the last six months of data collection, a computer program was made available in the surgical department that could be used to give diagnostic advice. The program, written by the author, was implemented in Microsoft BASIC on an APRICOT xi personal computer.

In order to obtain advice, the user typed in the code numbers of collected case details, from the AAP form. The program then accessed the AAP database (6) and used a simple Bayesian algorithm [3.1.a.i] to calculate the relative likelihood of each of 7 surgical emergencies and 2 non-surgical conditions being present. The system was set to offer a ranked differential list providing that any single disease likelihood score was greater than or equal to 50% of the total for all considered diseases. If no disease scored over 49% then the system would advise that there was insufficient data.

The computer also stored clinical information, to allow later input validation and offered a printed summary of case details and advice given, for inclusion in the case notes.

f) Explanation of System Output

Two explanation routines were incorporated into the computer system.

(i) Feature Ranking

Feature ranking could be used to identify the factors, from history and examination, which had proved most important in determining the calculated result. Symptoms and signs that had been entered were ranked according to the strength with which they supported or did not support a particular diagnosis. Ordering was made according by comparison of frequency information in the database. For example, when finding the features supported disease d1, rather than disease d2 the system would calculate;

$$\frac{P(sk|d1)}{P(sk|d1)+P(sk|d2)}$$

where $P(sk|d1)$ = likelihood of symptom k given disease d1
$P(sk|d2)$ = likelihood of symptom k given disease d2

for each symptom and rank the results in descending order.

The explanation routine allowed comparison of the computer diagnosis with the doctor diagnosis in order to show features that strongly supported one or the other.

(ii) Summary of Diagnostic Features and Management Information

Single pages of diagnostic and management advice were available for each of the computer's diagnostic end-points. These could be displayed once case information had been entered.

g) Consolidation Phase

A period of consolidation followed conclusion of the third phase of the trial. Cases were followed up if information was missing, the final diagnosis was in doubt (including all cases of NSAP) or where further relevant admissions occurred.

h) Analysis

Following consolidation the collected data were analysed with respect to the described surgical activity indicators according to phase of the trial. The reasons for system diagnostic misclassification have been assessed by a comparison of individual and grouped case data.

At the conclusion of their involvement in the trial, doctors were asked to comment on whether they had found the data collection sheet and computer system to be of value in clinical practice. They were also invited to record any particularly good or bad points about the way the system had been implemented.

3) Comparisons of the Performance of The Leeds Acute Abdominal Pain Diagnostic Advice System with Paramedics, Non-Medical Staff and Referring General Practitioners

The Haslar trial allowed the collection of a set of detailed standardized case summaries. For each case, several opinions of diagnosis were available, from the initial assessment made by the general practitioner, to that of the consultant following investigation. A clinically important subset of 'suspected appendicitis' cases has been used in comparisons of diagnostic accuracy between the computer, doctors and paramedics. Cases have been assigned to the 'suspected appendicitis' group if at least one examining doctor recorded appendicitis as being the likely cause of presenting symptoms and signs.

a) Comparison of the Diagnostic Performance of Computer and General Practitioner

The diagnosis assigned by computer and referring general practitioner has been compared for each of the 'suspected appendicitis' cases. A 2x2 table has been constructed and McNemar's test applied to discordant pairs.

b) Comparison of the Diagnostic Performance of Computer and Paramedic, Investigate Relevance of Paramedic Management Plans

Refresher medical training for seagoing RN paramedics is conducted at the Royal Navy Staff Training School. The members of several such courses were each given the (anonymous) history and examination details of 2 or 3 randomly distributed 'suspected appendicitis' cases. They were allowed to use reference books, normally supplied at sea and after one hour were asked to give their diagnosis and plan of management for each case.

The paramedics were motivated to participate by being informed that the results of their deliberations might be used as part of the course assessment. The staff running the trial were blinded to the diagnosis in each case, so that they would not inadvertently favour any of those being tested. No case discussion was allowed either between paramedics or between paramedics and staff.

Following completion of the tests, the management performance of paramedics was compared with that of the doctors who had originally managed the cases. Diagnostic accuracy was compared with that of the computer. A 2x2 table has been constructed and McNemar's test applied to discordant pairs.

c) Investigation of the Ability of Paramedical and Non-Medical Staff to Collect Clinical Information from Patients Suffering Acute Abdominal Pain

A trial was performed to investigate whether it was possible for RN personnel who were not medically qualified to collect sufficient accurate medical information from patients to allow use of the Leeds AAP advice system.

A simplified common language version of the AAP data collection form and associated guidance notes were developed (217) by me in conjunction with doctors in the surgical department at RNH Haslar.

Experienced seagoing coxswains, were trained during a one hour session to take a relevant history and carry out an abdominal examination. The trained coxswains placed themselves on a call-out list and were summoned from time to time to examine male service patients who had been admitted with AAP.

The opportunity was also taken during the trial to test paramedical staff, who used the data sheet and guidance book, but received no specific additional training.

The history and examination details collected by the coxswain were compared with a Leeds proforma filled in by the duty house officer and a prediction of diagnosis was made from both data sets using the Haslar AAP program.

4) The Design and Construction of DERMIS: A Primary Care Advice Dermatology Diagnostic Advice System

a) Investigation of Referral Patterns of General Practitioners Managing Patients with Skin Disease

A survey has been conducted of the reasons general practitioners give for referring patients to a dermatology clinic. Appropriate information has been extracted from referral letters found in the notes of consecutive cases appearing at the clinic. A summary table has been produced where the reason for referral has been compared with management outcome.

b) Prospective Collection of Clinical Records and DERMIS Database Formation

A multiple choice data collection sheet has been designed that incorporates most of the clinically descriptive terms used in dermatology. Each of the terms has been simply defined and the definition tested on students and general practitioner trainees. Definition testing and modification continued until each achieved a 'stable state' that was easily understood yet conveyed a single and appropriate meaning;

For example, the border of a lesion is described as being 'definite' if it is sufficiently well demarcated to allow a line to be drawn all around it.

The data collection sheet is self copying and is in two parts. One part becomes the patient record and the other which cross-refers clinical responses to a series of numeric codes has been used for transfer of the information to electronic media.

The subjects have included 5203 consecutive new patients referred to a dermatologist between 1985 and 1989. The diagnosis for each case was made clinically by the dermatologist at the time of examination, or if there was uncertainty, at a later date in the light of histological or mycological findings. Diagnosis was routinely checked in all cases attending for review.

In practice, data collection was performed by a number of observers, who were overseen by the dermatologist. Each case was coded twice and inconsistencies between the copies were eliminated by reference to the original. Extensive logic checks were applied to each case following coding in an attempt to detect data collection errors.

The frequencies with which the collected symptoms and signs occurred in diseases were estimated using cases in the database.

c) <u>The Calculation of Frequencies</u>

The collection of details about cases results in a mass of information which, as a whole, rapidly becomes difficult to manipulate and virtually impossible to assimilate.

In order to make the information intelligible a process of averaging is often used to generate frequencies of occurrence.

$$\begin{array}{c}\text{Estimated frequency } (f) \text{ of} \\ \text{occurrence of symptom } (s) = \\ \text{in disease } (d)\end{array} \qquad \frac{\text{Number (n) of cases of d with}}{\text{Number of cases of d}}$$

The frequency obtained can be expressed as a percentage and is often used as an estimate of probability.

is often used as an estimate of probability.

The standard error of a frequency estimate can be calculated by;

$$SE(f)= \sqrt{\frac{(f(100-f)}{n}}$$

d) The Use of a Bayesian Inference Algorithm in Diagnostic Prediction

The simple algorithm described in [3.1.a.i] was applied in initial tests of system diagnostic accuracy where the independence of symptoms and signs was assumed. Application of the algorithm produced a differential list ranked by relative likelihood score. Implementation has been in MUMPS on an IBM compatible desk top computer.

5) Investigation of Measures that Can be Taken to Improve the Performance of Diagnostic Advice Systems that Use a Simple Bayesian Model

The performance of laboratory tests in the three clinical domains has resulted in the collection of 3 clinical databases. These have been used to investigate known weaknesses and application problems encountered in the use of simple Bayesian predictive models.

a) Tests of the Assumption of Independence of Variables

First order feature association within each disease group has been investigated for the three clinical databases using a chi-squared test with a null hypothesis that the symptoms

The tests were performed automatically by computer. For each pair of features a 2x2 table would be created for the possible combinations of feature presence and absence [Figure 5];

Figure 5
Assessment of Association Between Variables

		symptom k'		
		present	absent	
symptom k	present	a	b	a + b
	absent	c	d	c + d
		a + c	b + d	N

where $$\chi^2 = \frac{(ad - bc) \times N}{(a + b)(c + d)(a + c)(b + d)}$$

with 1 degree of freedom

Positively and negatively associated variables ($p<0.05$) have been charted in order to assess the extent and clinical relevance of feature association within individual disease groups and within whole databases.

b) Taking Associations Between Variables into Account in Inference by Iterative Selection of Variables

One way of taking associations between variables into account in diagnostic inference is to eliminate redundant variables. The following iterative method is a development of one used by Teather (80) which has similarities to one reported by Goldman (89). The process was performed automatically by computer.

The mechanism of operation of the program was as follows;

- look at all the symptoms and signs that have not been used in this branch of the decision tree
- select the feature that produces the greatest differentiation between the two disease groups
- partition the groups and assess the significance of the differentiation achieved
- repeat within the branch until no further advantage is produced
- backtrack and test another branch

It has been applied to 2 clinical problems within the collected domain databases in order to produce diagnostic flow charts for differentiation between;

- appendicitis and other causes of 'suspected appendicitis'
- basal cell carcinoma (BCC) and solar keratosis (SK)

In the pair BCC and SK, the diagnostic flowchart was created using the first 270 cases of the two diseases appearing in the dermatology database. The chart was then used to predict diagnosis in the remaining 109 cases of the two diseases. The training set was also used to produce a database of frequency estimates for the two diseases. A simple Bayesian algorithm was then applied to the test set in order to predict diagnosis in each case. The diagnostic accuracy obtained by the two methods has been compared.

c) Taking Associations Between Variables into Account in Inference by Substitution of Combined Probability Estimates

Account can be taken of association between variables in prediction by treating combinations of associated variables as independent units.

A method of dynamic combined frequency estimate substitution has been devised which works as follows;

- the associations are found between pairs of variables within disease groups in a training set by Chi-squared analysis [3.5.a]
- the pairs are are ranked in descending order according to strength of association
- the list of known associations is sequentially compared with the features occurring in each test case. Feature combinations found in the test case are 'marked' if the individual features have not formed part of a previously 'marked' pair for that case
- a simple Bayesian model that assumes independence of variables is applied to test cases. Where 'marked' pairs occur, combined frequency estimates are used in prediction for all diseases in place of independent frequency estimates for the individual members i.e., the pair is treated as an independent variable

This method has been compared with a simple Bayesian model with regard to accuracy of differentiation between;

- appendicitis and other causes of 'suspected appendicitis'
- basal cell carcinoma (BCC) and solar keratosis (SK)

as in [3.5.b] above.

The process was then repeated for marked 'triplets'. Marked 'triplets' occurred where strong associations were found between each of three variables. On testing, combined 'triplet' frequency estimates were produced and substituted in preference to 'pairs' or single variable frequency estimates.

d) Production of a DERMIS Reduced Dataset by Elimination of Redundant Variables

The associations between the presence of collected symptoms and signs, and diseases have been estimated by chi-squared analysis [Figure 5].

During development of DERMIS at a point when 3508 cases had been collected, ranked lists of associations between features and diseases were produced. The lists were compared and features found to have little association with any disease submitted to the dermatologist for an opinion concerning their relevance to dermatology diagnosis.

Where no reasons could be found for retaining features they were removed from a list of features proposed for implementation in DERMIS. This list is referred to as the 'reduced dataset'.

e) Determination of the Number of End-Points to be Used in Prediction by DERMIS

By the time 2921 dermatology cases had been collected, 182 separate diagnoses had been identified. These end-points represented sub-groups of clinically identifiable families of diseases.

The database was split into a training set of 2538 cases and a test set of 383 cases. A frequency database was formed where all 182 diseases were considered separately. A simple Bayesian algorithm was used to predict diagnosis in the test set cases. Failures of prediction were investigated with respect to confusion occurring within disease families.

The total number of disease groups was then reduced by 'clinically appropriate' combination. 'Clinically appropriate' combination involved diseases that could be managed in the same way, where the dermatologist considered the result formed an acceptable referral grouping for use in general practice. A result of this process, was the development of a 32 clinical end-point group model, which was then evaluated using the training and test cases in Bayesian prediction. The effects of group reduction before and after prediction were also assessed. Further analysis of failures of prediction lead to the revision of grouping criteria and development of a 42 clinical end-point group model.

e) Selection of a Lower Frequency Bound Estimator

During the tests described in [3.5.d] above, the opportunity was taken to investigate the performance of three lower bound frequency estimators. This type of estimator is used when no information is available concerning the relationship between a particular feature and disease.

In the Leeds abdominal pain database, lower frequency bounds are all set to equal 0.1.

Three methods of estimating the lower bound were tested in sequence and the effects upon system diagnostic accuracy observed for the 182 disease group model. The three estimators applied when empty cells were found include;

(i) a probability of 0.1

(ii) a probability of 1/2n where n=number of cases in the disease group

(iii) A more complex estimator proposed by Perks (as described by Good) (218) that takes into account the number of options in each question as well as group size.

$$\text{Probability} \quad (S_j=k/D_i) \quad = \quad \frac{n_{ijk} + 1/c_j}{N_i + 1}$$

Where the j th symptom S_j is in the k th category, D_i is in the i th disease, n_{ijk} is the number of cases of D_i with symptom S_j in the k th category. N_i is the number of of cases of D_i and c_j is the number of categories for the j th symptom S_j.

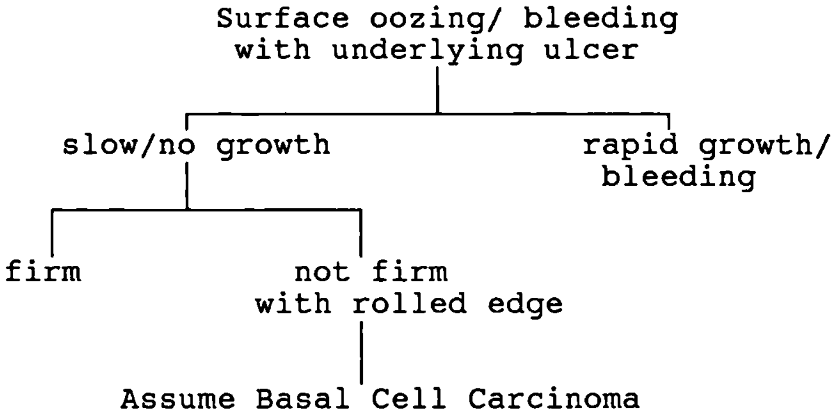
The probability of D_i is based on the estimate N_i/N (N is the number of cases for all diagnoses).

e) The Representation and Reliability of Expert Beliefs

Richard Ashton, the dermatologist involved with data collection for DERMIS, has independently developed and terms used are those that have been developed for DERMIS. The algorithms represent the dermatologist's beliefs concerning the importance of particular features and combinations of features in diagnosis.

The algorithms for identification of three common diseases, psoriasis, solar keratosis and basal cell carcinoma have been encoded for computer use and have been used to predict the presence or absence of the named diseases in the dermatology database of 5336 prospectively collected cases.

Figure 6
Ashton Algorithms: Extract from One of Several Pathways that Can Result in Prediction of Basal Cell Carcinoma


Surface oozing/ bleeding with underlying ulcer
slow/no growth
rapid growth/ bleeding
firm
not firm with rolled edge
Assume Basal Cell Carcinoma

The predictions produced by this method have been compared with predictions produced by a Bayesian model that has been derived from the database and tested by 'one out' analysis. The Bayesian model used 42 clinical end-point groups which included those covered by the coded algorithms (the DERMIS configuration was; 42 groups, reduced data set, frequency combination allowed, Perks estimator in force).

combination allowed, Perks estimator in force).

6) Laboratory Tests of the Performance of the DERMIS System

A series of tests of diagnostic prediction have been made using the dermatology database of 5203 cases and a simple Bayesian inference method (3.1.a.i). In each of the following evaluations, an iterative 'one out ' test method has been employed;

'One Out' Test Method

```
- a frequency database is formed using all of the cases
- cases are selected in order
  loop ....
        - a case is selected
        - the frequency database is adjusted to take
          account of absence of the presenting case
        - a prediction of diagnosis is made using the
          Bayesian algorithm
        - the frequency database is restored
      ....
```

i.e. each case is selected and compared with all of the remaining cases.

The following configurations have been compared with an initial configuration of the system that comprised 221 disease groups with a full data set and lower frequency bounds set by Perks's estimator (where cells contained no information);

a) 221 disease groups, the reduced data set [3.5,d], Perks estimator

b) 42 disease groups [3.5.e], reduced data set, Perks estimator

c) The Use of Combined Frequency Estimates

The investigation of the associations between features of diseases has been described [3.5.a][3.5.c]. Where these associations have been found to occur between pairs of possible answers to particular questions on the dermatology data collection sheet, a combined frequency estimate has been used in place of independent estimates when applying the predictive model.

d) Application of Expert Belief to Lower Bound Estimates

Following collection of the dermatology database it was found that many estimates of the frequency of occurrence of features within diseases fell below 5%. All such examples were referred to the dermatologist for an opinion as to whether the feature did or did not occur in the disease. Where the dermatologist was certain of non-feature occurrence, this was used in prediction as a means of eliminating diseases from the differential.

7) Trials of the DERMIS System in Clinical Practice

a) The Advice Required by General Practitioners Compared with the Advice Available from DERMIS

A further survey of the referral habits of general practitioners has been conducted amongst 125 cases randomly selected from the dermatology database. In each case appropriate information has been extracted from general practitioner's referral letter. Summary tables been compiled where the reason for referral has been compared with management outcome and DERMIS advice (based on 'one out' calculation).

management outcome and DERMIS advice (based on 'one out' calculation).

b) The Choice of User Interface and Explanation Routines

Various development models of DERMIS have been made available to doctors and students working in the dermatology clinic in order to allow live testing of user-interfaces and explanation routines.

The three methods of entering data tested were as follows;

(i) Doctor completes a data collection sheet then enters numerical or mnemonic codes into the computer, via the keyboard.

(ii) Doctor completes a data collection sheet that rests on a touch sensitive input device. Ticks are sensed by the device and codes automatically entered into the computer.

(iii) Doctor uses a keyboard to select appropriate answers from menus using single key presses.

Various explanation routines have been provided, term definition ;

(i) The user is able to add data to, and subtract data from, the case record in any order. He is presented with an immediate update of the relative likelihood output.

(ii) Production of ordered lists of features that support any selected diagnosis rather than any other, or rather than all the rest. The method employed was a development of that described at

(iii) Production of a ranked list of features that are critical to the order of the current system differential. This is produced by an iterative process, conducted for each feature entered, of recalculating the differential with the postulate that the feature has been removed. The result is produced in less that 5 seconds on an IBM compatible 286 portable running MUMPS.

(iv) Production of a list of diseases excluded by expert opinion and features that have caused exclusion.

The data entry and explanation routines described have been subjected to extensive testing over a period of several years. There has been measurement of entry times using particular methods, usage of explanation routines and usage of the system itself. There has also been subjective assessment of the diagnostic performance of the system by users.

c) Semi-Field Trial of Dermis as a Decision Support Tool for Primary Care

Photographs are regularly taken during dermatology clinics at Haslar Hospital as part of patient work up. 25 recent, case records of fully worked up patients, were randomly selected from amongst those that contained photographs. The dermatologist then chose 8 of these as being representative of the range of common referrals to the dermatology clinic.

None of the cases had been included in the DERMIS database or used in previous testing. They were selected without reference to the DERMIS system. The dermatologist viewed each of the sets of photographs in order to check that salient features of lesions and rashes had been adequately reproduced.

salient features of lesions and rashes had been adequately reproduced.
Trials took place between 1991 and 1992, in three locations and involved 49 general practitioners, and 9 hospital doctors who were not trained dermatologists.

Each doctor was presented with a handout that contained the definitions of terms used in DERMIS. The definitions were then reviewed during a 15 minute training session. Results of laboratory tests of the diagnostic performance of the DERMIS system were also described.Cases were presented to the doctors sequentially. Doctors were encouraged not to confer.

A partially completed data collection form was provided for each test case. The information supplied was that which could not be obtained from viewing the case photographs, such as age, duration of the problem, previous history, etc.

Photographs of the dermatological lesions associated with each case were shown to the doctors and they were asked to record their findings on the appropriate data collection sheet. They were then asked to decide their diagnosis and record both that and a brief management plan.

A representative selection of possible outputs from the DERMIS system were then described. These reflected different combinations of findings that might have been collected by the doctors. Doctors were able to request individual advice if their assessment had not been covered. Doctors were then asked to write down a diagnosis and management plan made in the light of the computer's advice.

Once all the test cases had been presented and data collection sheets returned, the 'true' diagnoses were revealed.

For each case, the data items collected by participating doctors have been compared with those recorded by the dermatologist. The decisions concerning diagnosis and management of cases made before receipt of DERMIS advice have been compared with those made later. Comparisons have also been made with 'true' diagnosis and the dermatologist's recommended management.

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Chapter 4

Results Arising From Experimental Work Performed

The results obtained from the work carried out for this thesis are now described. The order of presentation follows that adopted for both the plan of work and description of methods.

1) Comparison of Inference Models for Acute Chest Pain Diagnosis

The published details and performance of a number of the acute CP decision support systems discussed in this thesis are summarised in [Table 3]

Table 3

Summary of Inference Methods Used and Reported Performance of Several Acute Chest Pain Advice Systems

Author	Variables	Method	Subjects	Decision	Sens%	Spec%
deDombal	History Examination ECG Enzymes	Bayes	973 Presenting to GPs	MI Not MI	94.7	95
When Tested by Goldman			<900 presenting to hospital		52	87
Goldman vers. (i)	History Examination ECG Opinion	Tree	900 presenting casualty /admitted	MI Not MI	90	65
When Tested by Poretsky			168 suspected MI		80	62
Goldman vers.(ii)			4770 in casualty		88	74
Joswig	History Examination ECG Biochem	Logist Regres	173 prior to Angio-graphy	Coronary artery changes	88	84
Pozen	History ECG	Logist Regres	2320 presenting to hospital	MI Not MI	98	96
Wyatt	History Examination ECG (auto)	250+ Rules	150 presenting to hospital	CCU See Soon Wait	88	80

Table 3
Summary of Inference Methods Used and Reported Performance of Several Acute Chest Pain Advice Systems (continued)

				Decision	Sens%	Spec%
CLINICIANS						
by de Dombal					<95	<95
Poretsky				MI /	80	85
Goldman				Not MI	87.8	71
Wyatt					88	93

b) Subjects and Data Collection

Data collection forms were used during admission clerking of 108 ACP patients that were admitted to the CCU. Four forms, were not completed because of the poor condition of the patient on arrival. The remaining forms had few items of missing data. In all cases, missing information could be obtained from other hospital records.

The age/sex distribution of the case set is given in

[Table 4]

Table 4
104 CCU Admissions:
Age / Sex Distribution of Patients Included

Age	Male	Female
30-39	3	0
40-49	11	0
50-59	22	3
60-69	32	8
70-79	15	6
80+	4	0
Total	87	17

c) Diagnostic Classification

The distribution of cases according to final diagnosis, is given in
[Table 5].

Table 5
104 CCU Admissions: Final Diagnosis

Disease Group	Number
AMI	71
Angina	22
Other	11
Total	104

Two of the included patients were readmitted during the study for separate episodes of acute chest pain. Amongst the group as a whole, 44 (42%) had a history of previous AMI or angina. Five of those who had suffered previous AMI also suffered with angina. A summary of previous cardiac events in the study group is given in [Table 6].

Table 6
104 CCU Admissions:
Previous Cardiac Events in Test Group

		History Prior To Admission	
		AMI	ANGINA
Final	AMI	13	11
Diagnosis	Angina	5	12
(this admission)	Other	2	1
	Total	20	24

d) System Comparisons

Comparisons of classification efficiency have been made between the ACP advice models. In [Table 7], model sensitivity, specificity and accuracy levels, for AMI identification, are compared with the accuracy of the admitting physicians and that of a cardiology registrar's assessment of an initial 12 lead ECG.

Table 7
Sensitivity and Specificity for Models Considered When Tested on 104 CCU Admissions: Distinction of AMI from Not AMI.

Model	Sensitivity	Specificity	Accuracy
Pozen	67	53	64
Goldman (i)	89	57	80
(ii)	96	50	82
de Dombal	92	47	79
Clinicians*	*	*	68*
ECG **	68	90	73

* Concerns admission from casualty

** 12 Lead ECG taken shortly after admission and read by cardiology registrar

Predictions of a cardiac cause for ACP could be obtained from three of the models. Their accuracy, in this mode, has been compared with that of the admitting physicians and a cardiology registrar's assessment of an initial 12 lead ECG, in [Table 8]

Table 8
Sensitivity and Specificity for Models Considered When Tested on 104 CCU Admissions: Distinction of Cardiac From Not Cardiac

Model	Sensitivity	Specificity	Accuracy
Pozen	62	18	61
Joswig	81	13	76
de Dombal	99	63	96
Clinicians *	*	*	93
ECG **	86	50	83

In [Table 9] comparisons of accuracy in prediction of acute MI are made between the advice systems. The results of applying McNemar's test to the differences found are also shown. The systems produced by both Goldman and de Dombal have been found to classify the ACP cases with significantly greater accuracy than the system produced by Pozen. These two advisers also appeared to show greater accuracy of diagnostic group assignment than unaided admitting physicians ($p<0.05 >0.01$).

Table 9

Comparison of the Accuracy of Acute MI Prediction Between Advice Systems Tested on 104 CCU Admissions: McNemar's Test Applied to Differences

Advice System	« B »			
A	Pozen	deDombal	Clinicians	ECG
Goldman (i)	++	0	+	0
Goldman (ii)	++	0	+	0
de Dombal	+		+	0
Clinicians	0			
ECG	0		0	

Where;
$v = 1$
++ = Advice system A accuracy found greater than B, $p < 0.01$
+ = Advice system A accuracy found greater than B, $p < 0.05$
0 = Advice system A accuracy found greater, but difference not significant

Comparisons of accuracy in prediction of the presence of a cardiac cause for acute CP have also been made. McNemar's test has been applied to the differences found. The results are given in [Table 10]. The system produced by de Dombal has been found to classify the acute CP cases with significantly greater accuracy than systems produced by both Pozen and Joswig.

Table 10
Comparison of the Accuracy of Prediction of a Cardiac Cause for Acute Chest Pain Between Advice Systems Tested on 104 CCU Admissions: McNemar's Test Applied to Differences

Advice System		« B »	
A	Joswig	Pozen	ECG
de Dombal	+++	+++	++
ECG	0	++	

Where;

v = 1

+++ = Advice system A accuracy found greater than B, p < 0.001

++ = Advice system A accuracy found greater than B, P < 0.01

+ = Advice system A accuracy found greater than B, P < 0.05

0 = Advice system A accuracy found greater, but difference not significant

2) <u>Hospital Trial of The Leeds Acute Abdominal Pain Diagnostic Advice System</u>

a) <u>Subjects</u>

A total of 353 patients have been included in the Hospital trial. Their distribution by diagnosis and phase of trial at time of presentation is given in [Table 11], where a comparison is made with the findings of a 1982 OMGE survey and the 1983-85 National AAP trial (9). The Haslar intervention phase columns in the following tables indicate summary statistics for both the 'data collection sheet only' and 'computer access' phases of the trial.

Table 11

Comparison of Hospital Admission Rates for Diseases Causing Acute Abdominal Pain found in National AAP trial, OMGE survey and Haslar AAP study.

Final Diagnosis	OMGE Survey 1982 n= 8480	National Trial -1986 n= 16737	Haslar Baseline 1 n= 167	Haslar Intervention 2 n= 186
	%	%	%	%
NSAP	36.1	58.6	42.5	43.5
Appendicitis	27.6	12.4	20.9	32.2
Cholecystitis	9.4	4.0	3.6	6.5
Gynaecology	3.3	1.9	4.1	4.3
Renal Colic	3.1	3.0	2.9	2.1
Pancreatitis	2.7	1.4	3.6	1.6
Perf Ulcer	2.7	1.3	1.8	1.1
Other	15.1	17.4	12.6	8.7

b) Data Collection

Two surgical firms work in Haslar hospital. Normally, one middle grade surgeon and one house officer (HO) are attached to each firm and house officer appointments last six months. During the baseline phase of the trial, house officer appointing was delayed by one month. Accordingly, it was decided to extend this phase to last seven months. The second phase of the trial was reduced by one month as house officers were only in post for five months. In the third phase of the trial, one house officer left after three months. The remaining house officer was required to work for both firms and ceased to collect information or use the computer. The third phase has therefore been considered as having lasted three months. For the purpose of comparison, the middle grade surgeons have been classed as senior house officers (SHO).

c) Performance of the Doctors and Surgical Firms

The cases collected were compared with those registered on the hospital patient administration system. During the baseline part of the trial, information was obtained about all AAP admissions. It is estimated that forms were completed for only 90% of admissions during the 'forms only' and 'computer access' phases of the trial. Of the 70 forms completed whilst the computer program was available, only 39 were used to obtain an advisory print out at the time of admission.

When the results of the 'forms only' and 'computer access' phases of the trial were compared, no significant differences were found between the overall performance rates of the doctors or the surgical firms. Accordingly, the data from both of these phases have been combined and reported as a single 'intervention' phase.

A comparison, by phase, of the measured surgical department activity indicators has been made made in [Table 12].

Table 12
Haslar Trial of Acute Abdominal Pain Advice System:
Surgical Unit Performance

Indicator	Haslar		National	
	Base %	Interv %	Base %	Interv %
HO Accuracy	49+	65+	46	65
SHO Accuracy	64+	79+	58	74
Neg Lap	14.6˜	7.6˜	16.4	10.0
Perf APPX	14.3	13.3	23.7	11.5

Where;

+ Difference is significant ($p < 0.001$) by SND test
˜ Difference is significant ($p < 0.05$) by SND test

Neg Lap = Negative laparotomy, the rate is % of total
laparotomies for appendicitis
Perf APPX = Perforated appendix,
the rate is % of total cases of appendicitis.

The recorded diagnostic accuracy of both junior and senior house officers dealing with AAP cases rose from 49% and 64% (+) respectively during the baseline period to 65% and 79% (+) during intervention. There was also a fall in the negative laparotomy rate during the second part of the trial(˜) and slight reductions in the perforation rate and stay time.

In [Table 13] the changes in diagnostic accuracy have been related to disease groups.

Table 13
Haslar Trial of Acute Abdominal Pain Advice System:
Diagnostic accuracy % by diagnosis

Final Diagnosis	Haslar Medical Officer Baseline %		Intervention %	
	HO	SHO	HO	SHO
NSAP	45	66	48	74
Appendicitis	68	82	85	93
Cholecystitis	(4/6)	(6/6)	(7/12)	(9/12)
Gynaecology	(1/7)	(2/7)	(6/8)	(4/8)
Renal Colic	(1/5)	(3/5)	(4/4)	(4/4)
Pancreatitis	(1/6)	(4/6)	(2/3)	(2/3)
Perf Ulcer	(2/3)	(3/3)	(0/2)	(0/2)
Other	62	71	75	81
Overall	49	64	65	79

Numbers correct and group size (/) are given, where total sub-set is small.

In both baseline and intervention phases of the trial, the majority of admissions were due to NSAP and appendicitis (63.4% & 75.7%). For both HOs and SHOs The greatest improvement in accuracy occurred in the diagnosis of appendicitis

d) Performance of The Computer

The computer advice system was used on 39 occasions and exclusively by house officers. On 15 of these occasions, a print out of the computer's prediction was filed somewhere in the patient's record. In the remaining 24 cases, the computer's advice was written down as part of the clerking process.

Of the 39 cases presented to the computer, 36 were examples of 'suspected appendicitis'. The computer was able to correctly identify the diagnosis of 26 of these cases. Computer advice was sought for a further 3 'difficult' clinical presentations of AAP. Here the computer's output was of little value because the diseases suffered did not appear in the database.

e) Explanation Routines

House officers made no use of the computer routines that had been designed to give further information about the diagnostic conclusions reached.

f) Users' Opinion

Four house officers used data collection sheets. They all found the sheets to be quick and easy to use. One house officer regularly included a completed data collection sheet in the patient record in place of his abdominal clerking, but the other three duplicated at least part, if not all, of the information as written notes. They all agreed that use of the forms increased the clerking time. Data collection sheets were not completed at times when house officers were busy.

One of the two house officers who had access to the computer accounted for 32 of the total number of cases entered. The computer was located in the side office to a ward and it was reported that approximately 5 minutes were spent in the room on each occasion that advice was sought. Both house officers complained that the siting of the computer was inappropriate for their pattern of work as they would often be required to clerk patients on a different ward. For this reason, in practice, they both tended to delay the entry of data into the computer until it was convenient for them to do so. The computer was not used when house officers were particularly busy. Their overall impression was that any advantages offered by the computer were outweighed by the time penalty incurred in its use.

3) Comparisons of the Performance of The Leeds Acute Abdominal Pain Diagnostic Advice System with Paramedics, Non-Medical Staff and Referring General Practitioners

At Haslar, emergency referrals from GPs and Establishments are admitted directly to the wards, without being seen in the casualty department. The accuracy of the referral diagnosis during the trial is considered in [Table 14].

Table 14
Haslar Trial of Acute Abdominal Pain Advice System:
Accuracy of referral diagnosis in 249 cases (70.5% of total)

Final Diagnosis	Referred Number	Referral Diagnosis correct %
NSAP	108	17.6%
Appendicitis	76	76.3%
Cholecystitis	19	47.4%
Gynaecology	6	(0/6)
Renal Colic	7	(4/7)
Pancreatitis	9	(1/9)
Perf Ulcer	2	(1/2)
Other	22	68.2%

Although sufferers of NSAP formed the largest admission group, very few (17.6%) had been assigned the correct diagnostic label at the time of referral.

At the end of the trial, a computer prediction was obtained for each of the completed AAP proforma. The largest sub-group of this database consists of 99 cases of 'suspected appendicitis'. These cases have been used to compare the diagnostic accuracy of medical officers, paramedics and the computer. The distribution by final diagnosis is given in [Table 15].

Table 15
99 Suspected Appendicitis Cases:
Breakdown by Final Diagnosis

Final Diagnosis	Number of Cases
NSAP	42
Appendicitis	51
Other Surgical	2
Gynaecology	4
Total	99

The accuracy of the computer in diagnosing patients with 'suspected appendicitis' has been compared with the relative likelihood output score produced for the correct diagnosis by using linear regression. The variance accounted for was 70.2%, b=1.0613, constant=11.2, v=4, $p < 0.05$. By this method, the predicted accuracy of the computer is 83.7% when producing a relative likelihood score of 100.

a) Comparison of the Diagnostic Performance of Computer and General Practitioner

The accuracy of the referral diagnosis of general practitioners managing cases of 'suspected appendicitis' has been compared with the accuracy of the computer in Table 16.

Table 16

99 Suspected Appendicitis Cases, Use of Computer:
Comparison of Accuracy of Assignment to Diagnostic Group
Between General Practitioners and Computer

	General Practitioner Assignment to Diagnostic Group		
	Correct	Incorrect	Tot
Computer Correct	44	30	74
Computer Incorrect	4	21	25
Total	48	51	99

GP correct = 48%
Computer correct = 74%
Significance of difference $p < 0.001$ (McNemar)

Computer advice reasonable in 30 cases
Computer of no value in 4 cases

On 34 occasions, the advice of the computer differed from that of the GP. In 30 cases the computer was correct, but in the remaining 4 cases, the GP produced the correct diagnostic classification. The computer missed 8 cases of appendicitis that had been referred by general practitioners for specialist opinion.

b) Comparison of the Diagnostic Performance of Computer and Paramedic, Investigate Relevance of Paramedic Management Plans

(i) Problem Definition

Although it has been shown that AAP is one of the main causes of evacuation from submarines in the USN (1), similar statistics are not available for the RN. It is known, however, that more than 50 personnel per year are evacuated from warships at sea for urgent medical reasons and return to the UK for treatment at one of the military hospitals. Many others are landed for treatment at local hospitals and medical centres. As part of the process problem definition, the annual incidence rate of appendicitis has been estimated for the RN population by standardization, using UK national age specific incidence rates [Table 17].

Table 17
The Estimated Annual Incidence of Appendicitis by Age Group For UK National and Royal Navy Populations

Age	RN Personnel in group	Annual UK Appendicitis Cases per 1000	Annual RN Appendicitis Cases
17-19	2040	5.2	10.6
20-29	11637	1.5	18.0
30-39	4251	0.7	3.0
40-49	695	0.3	0.3
Estimated total number of cases per year			31.9

In UK males who attend hospital, NSAP is found to be the cause of AAP 2.18 times as frequently as appendicitis. In the RN population, therefore we might expect 100 or so cases of AAP to present each year that are severe enough to be considered candidates for hospital admission. This is likely to be a subset of a much larger group of all patients that present with abdominal pain.

The estimation of the risk of a case of AAP causing a medical evacuation at sea requires additional information, such as the proportion of time spent by vessels at sea, the male/female mix of the crew and the expected performance of the medical team.

(ii) Performance Test

When 40 sea-going paramedics, undergoing refresher training, were each given two or three summaries of the house officers findings in 'suspected appendicitis' cases, their overall diagnostic accuracy for the test set was found to be 48%. [Table 18] gives the agreement and disagreement between paramedic opinion and computer advice for the 99 cases used.

Table 18
99 Suspected Appendicitis Cases, Use of Computer: Comparison of Accuracy of Assignment to Diagnostic Group Between Paramedics and Computer

	Paramedic Assignment to Diagnostic Group		
	Correct	Incorrect	Tot
Computer Correct	52	22	74
Computer Incorrect	6	19	25
Total	58	41	99

Paramedic correct = 48%   Computer correct = 74%

Significance of difference p < 0.01 (McNemar)

Computer Advice Reasonable in 22 cases

Computer advice of no value in 6 cases

There were 28 cases where the advice of the computer differed from the paramedic's opinion. In 22 cases the computer was correct, but in the remaining 6 cases, of which one was acute appendicitis, the paramedic produced the correct diagnostic classification. The computer correctly identified 5 cases of appendicitis that the paramedics planned not to evacuate.

c) Investigation of the Ability of Paramedical and Non-Medical Staff to Collect Clinical Information from Patients Suffering Acute Abdominal Pain

In a test of the data collection skills of paramedics and non-medically trained personnel, 7 coxswains and 5 paramedics used a specially prepared data collection form when collecting information from patients suffering with AAP. [Table 19] gives a comparative summary of the cases studied.

In the following Table, the first number in the 'data items' column is the number of symptoms and signs ticked on the data sheet by the coxswain / paramedic.
The number in brackets (), following, is the number of differences in positive findings made by a house officer examining the same patient.

Table 19

Medical Data Collection

By Paramedical and Non-Medical (Coxswains) Personnel:

Summary of cases collected (15 cases)

Collected By	Data Items (Diffs)	Diagnosis		Where used
		computer	final	
Coxswain	31 (1)	APPX	APPX	Hospital
Coxswain	31 (5)	APPX	APPX	Hospital
Coxswain	30 (2)	NSAP	NSAP	Hospital
Coxswain	31 (2)	NSAP	NSAP	Hospital
Coxswain	31 (10)	APPX	NSAP	Hospital
Coxswain	31 (3)	NSAP	NSAP	Hospital
Coxswain	32 (4)	NSAP	NSAP	Hospital
Coxswain	32 (?)	APPX	NSAP	At Sea*
Coxswain	32 (?)	R Col	R Col	At Sea
Coxswain	32 (?)	R Col	R Col	At Sea
Paramedic	32 (0)	NSAP	NSAP	Sick Bay
Paramedic	32 (2)	NSAP	NSAP	Sick Bay
Paramedic	32 (1)	NSAP	NSAP	Sick Bay
Paramedic	32 (2)	R Col	R Col	Sick Bay
Paramedic	32 (3)	SMBOBS	SMBOBS	Sick Bay

Where;

APPX	= appendicitis
R Col	= Renal colic
SMBOBS	= Small bowel obstruction

*Negative Laparotomy performed Glasgow

Seven cases of AAP were seen by coxswains at Haslar and fully documented. The complete details of three further cases were forwarded by signal from submarines at sea. One of these patients was evacuated because of 'suspected appendicitis', operated upon, and subsequently found not to have had the disease.

In 6 of the 7 AAP cases dealt with by non-medical staff ashore, the computer was able to produce the correct diagnosis from the case details collected by both the non-medic and the examining house officer. In six patients seen by paramedics, both paramedics and reviewing doctors collected data that were sufficiently similar for the computer to produce the same (correct) diagnosis from each set.

d) <u>Comparison of Doctor, Paramedic and Computer Diagnostic Accuracy Rates When Dealing with Cases of 'Suspected Appendicitis'</u>

The accuracy of diagnostic classification by the computer program, doctors and paramedics is compared in [Table 20]. McNemar's test has been applied to the differences found.

Table 20

99 Suspected Appendicitis Cases, Performance of Practitioners and Computer Program:

Comparison of Accuracy of Assignment to Diagnostic Group

Assessment	Accur	Compared with B			
Made by A	%	GP*	HO	Comp	Para
HO	65	+			
SHO	79	+++	+	0	++
Computer	74	+++	0		++
Paramedic*	58	+			

Where;

v = 1

+++ = Diagnostic accuracy of A found greater than B, p < 0.001

++ = Diagnostic accuracy of A found greater than B, p < 0.01

+ = Diagnostic accuracy of A found greater than B, p < 0.05

0 = Diagnostic accuracy of A found greater, but difference not significant

** GP accuracy (=48%) assessed from analysis of referral letters

* Paramedic accuracy assessed from performance on case history information collected by HO

The SHO was, in general, able to produce significantly greater accuracy of classification than the other advisers considered, apart from the computer. The computer, using house officer data, proved to be more accurate in diagnosis than HOs, GPs and paramedics.

4) The Design and Construction of DERMIS: A Primary Care Dermatology Diagnostic Advice System

a) Investigation of Referral Patterns of General Practitioners Managing Patients with Skin Disease

In a survey of the reasons general practitioners gave for referring patients to the dermatology clinic,[Table 21] it was found that on 68% of occasions both diagnostic and management advice were required. In 55% of these cases the specialist recommended mangement that could have been provided by the general practitioner. A further 37% of patients referred for diagnosis and management had benign tumours removed.

Table 21
Comparison of General Practitioner's Reason for Referral with Specialist Advice Given in 211 Consecutive Cases Attending the Dermatology Clinic at Haslar Hospital;
Reason For Referral

		Diagnosis & Management	Further Management
Management	Routine Treatment Given	78 *	7
	Specialist Treatment Given	65 ˜	61
	Totals	143	68

Where;

* includes 70 cases in which final diagnostic group matched one of the 39 main DERMIS groups

˜ includes 24 cases in which a benign tumour was diagnosed then removed.

(ii) Application of Gold Standards

The 'Gold Standard' applied for diagnostic end-points has involved pathological sample analysis. However, samples were only taken when the dermatologist was any doubt about diagnosis. Cases have been routinely followed up to test whether such judgements are reliable. In one sample of 200 database cases, 8 % of clinic diagnoses had been changed following tissue sampling or other information available at susequent review. A further 28 cases, out of 5203 are now known to have been assigned an incorrect initial diagnosis.

5) Investigation of Measures that Can be Taken to Improve the Performance of Diagnostic Advice Systems that Use a Simple Bayesian Model

a) Tests of the Assumption of Independence of Variables

Examples of first order feature association within disease groups of the gathered clinical databases are given in [Figures 8,9,10,11]. In the figures, the first feature in each line has been found to be associated with each of the features, in brackets, that follows it.

FIGURE 7 Full Dermatology Data Collection Sheet

Date /___/ Completed by HOSPITAL No.

Male/ Female AGE. yrs/ mths NAME.

HISTORY

1 TIME since onset of rash/lesion years/ mths/ days
DURATION of rash when present > 8 weeks / weeks/ days/___hours
SIZE gradual increase in Yes/No PIGMENT any change in Yes/No
PETS dog/ cat/ other ( )

2 HISTORY of
atopic eczema family
other eczema family
asthma family
hay fever family
psoriasis family
dandruff
skin disease family
( )

3 ITCHING
none
mild
moderate
severe
day
night
fami y

4 RACE
caucasian
negro
asian
oriental
SKIN TYPE
I III
II IV

5 TREATMENT used previously
steroids improved
moisturisers improved
anti-fungals improved

6 MADE WORSE BY
soap detergent
oil grease
water
sunshine
exercise/hot baths

7 SKIN HYDRATION
normal
greasy
dry
very dry (ichthyosis)

OINTMENTS/CREAMS given by GP

DRUGS taken in last 3 months

OCCUPATION

EXAMINATION OF RASH/LESION

8 NUMBER
single lesion
2 5 lesions
6 20 lesions
21+/rash
no rash

9 *DISTRIBUTION*
symmetrica
asymmet ica
g ouped
linear
sun exp ed

10 SIZE
mm avera e
var ab e
(mm

11 TYPE
ma u e
patch
papule
nodu e
p q e
ve e
b a
p t e
we
e
e
te
a
e e
oedem
ne

12 COLOUR
normal
pink
red
p nk purp e
wh te
cream
orange
yellow
golden
ight brown
dark brow
bla k
g ev
hyp pigme ted
hyperpigmented
other
)

HOSPITAL N

13 BORDER E GE
def nite
variable
indistinct
raised ab ve entre
active edge
other
(

14 SHAPE def te e
rou d
ova
annular
linear
peduncu ated
irregu ar
other
(

RFA E FEA RES
n
warty
scaly
exudate
cru t
ma e ated
l ab e
tr pl
l at & h ny
wh te streak
he if ed
ex o ated
umb cated
na y
ke ol
t c
the
)

16 VASCULAR FEA
erythema
purpura
telangiecta a
varicose vei

17 PALPATION
(deep (surfa e
norma sm th
s ft u ev
f rm r u
hard h
tender

18 SCRATCH TE T
n hange
m ld s a e
profuse ca e
weal

ASSOCIATED FEATURES

19 SCALP
not involved
papules
scaling
hair loss
uniform
patchy
palpable
impalpable
scarring
extends beyond hair margin
remains within hair margin

20 MOUTH
not involved
white streaks
ulcers

21 GENITALS
not involved
involved

22 NAILS

f	t	not involved
f	t	*fine pitting*
f	t	coarse pitting
f	t	onycholysis
f	t	*subungual hyperkeratosis*
f	t	nail thickening
f	t	loss of nail plate
f	t	transverse ridges
f	t	longitudinal ridges

23 PALMS & SOLES
not involved
vesicles
pustules
fissures
hyperkeratosis
scaling in creases
scaling in finger webs
maceration between toes
burrows on fingers/wrists

MYCOLOGY PERFORMED YES/NO RESULT +ve/ ve
BIOPSY PERFORMED YES NO organism

DIAGNOSIS:

TREATMENT:

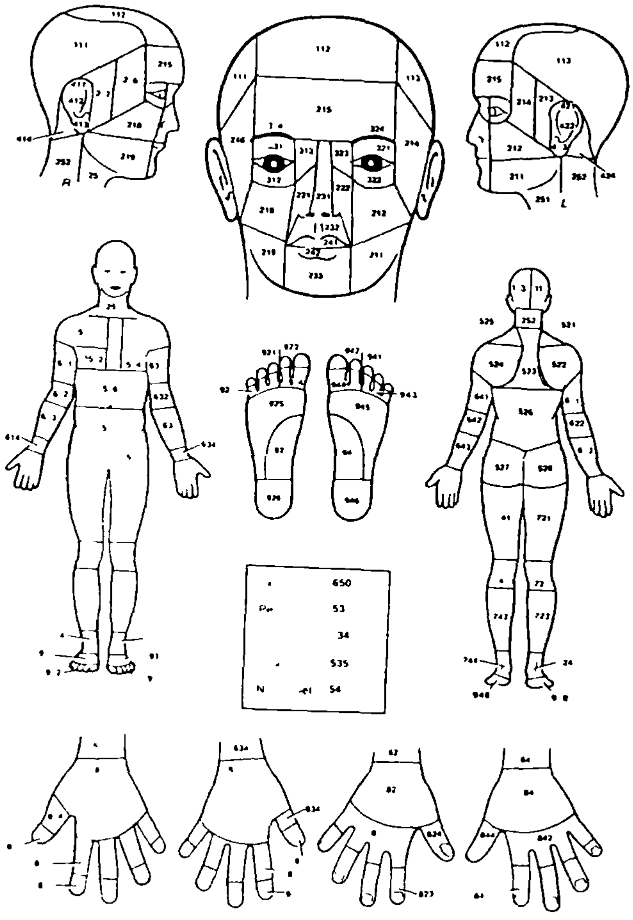


Figure 8

Appendicitis; examples of first order associated items

abdominal scar [abdominal surgery]

pain onset RLQ [flushed , aggrav by movement , pain steady]

pain onset central [pale , no anorexia , not flushed , no rectal tenderness]

pain now RLQ [movement poor]

movement poor [tender RLQ]

general tenderness [rebound , guarding , rectal tenderness]

progress worse [duration 12-23 hours]

duration to 12 hours [flushed , no rectal tenderness , progress same]

Figure 9

Myocardial Infarction: Examples of First Order Associated Terms

pain upper half chest	[getting better , relief diamorph]
relief diamorph	[pain upper half , severe pain , getting better , no nausea]
duration <6 hours	[steady pain , nausea , sweating]
duration 24 hours	[intermittent pain]
crushing pain	[severe pain , vomited , ST change]
no aggrav factors	[onset sudden , steady pain]
no relief	[breathless sitting]
breathless sitting	[pain central , remains same , no relief , creps heard]
colour pale	[fast respiration , distressed , cold clammy]
no oedema	[duration <6 hours]
sweating	[duration <6 hours , nausea , anxious , creps heard]
no sweating	[getting better , no nausea , mood normal]

Figure 10

Eczema: Examples of First Order Associated Terms

severe itch [excoriated ,lichenification]

exudate [crust]

erythema [pink , red]

Figure 11

Basal Cell Carcinoma: Examples of First Order Associated Terms

size 1-9 mm [papule , no size change , normal surface , raised edge]

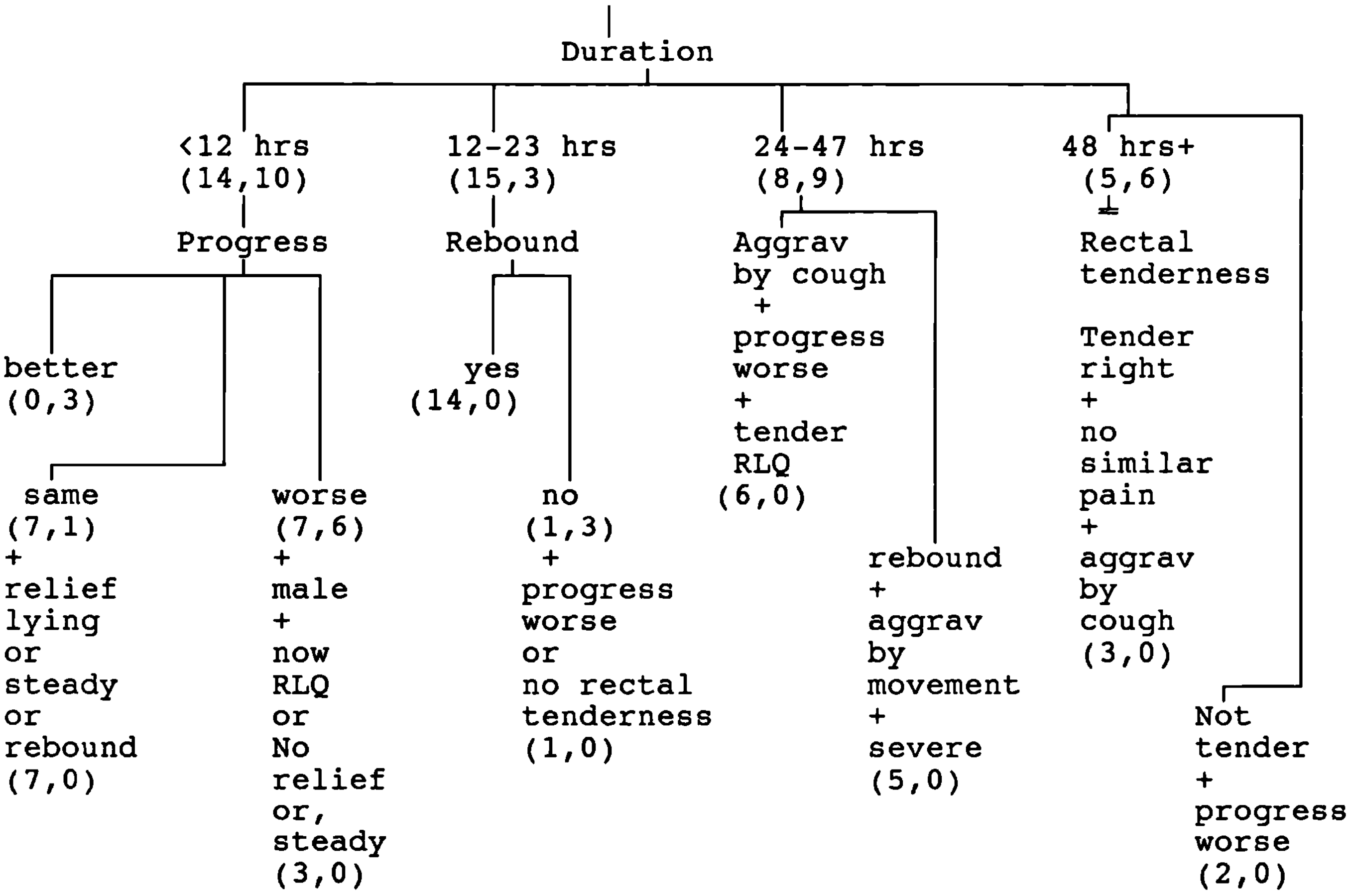
size 10-19 mm [nodule , size change , crust surface , ulcer]

b) <u>Taking Associations Between Variables into Account in Inference by Iterative Selection of Variables</u>

Two diagnostic flowcharts have been produced by an iterative partitioning method [3.5.b] from the 'suspected appendicitis' cases and the dermatology case database.

(i) Figure 12

Diagnostic Flowchart for the Differentiation of Appendicitis from Other Causes of Acute Abdominal Pain


Duration
<12 hrs (14,10)
Progress
better (0,3)
same (7,1) + relief lying or steady or rebound (7,0)
worse (7,6) + male + now RLQ or No relief or, steady (3,0)
12-23 hrs (15,3)
Rebound
yes (14,0)
no (1,3) + progress worse or no rectal tenderness (1,0)
24-47 hrs (8,9)
Aggrav by cough + progress worse + tender RLQ (6,0)
rebound + aggrav by movement + severe (5,0)
48 hrs+ (5,6)
Rectal tenderness
Tender right + no similar pain + aggrav by cough (3,0)
Not tender + progress worse (2,0)

Where the figures in brackets represent the split; (appendicitis , other cause).

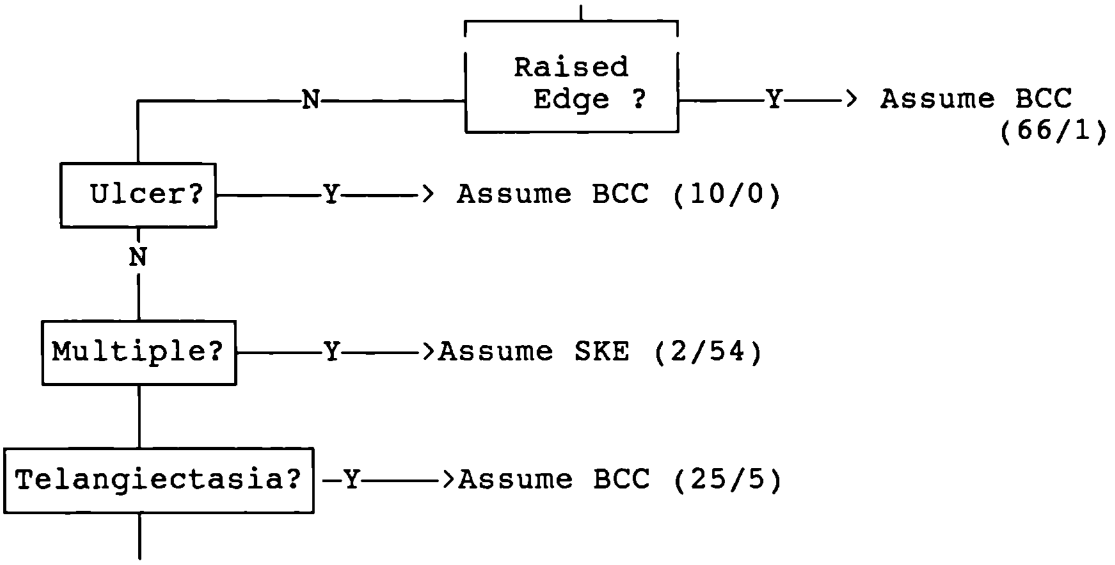
The tree is followed by asking sequential questions. For example;

Question		Example Answers
what is the duration of pain	?	less than 12 hours
is the progress, better/ same/ worse	?	same
is the pain steady	?	yes

Then predict appendicitis

(ii) Figure 13

First Part of a Diagnostic Flowchart for the Differentiation of Basal Cell Carcinoma from Solar Keratosis



Where the figures in brackets represent the split; (BCC/SKE)

(iii) <u>Comparison of a Diagnostic Flow Chart for Differentiating Basal Cell Carcinoma from Solar Keratosis with a Simple Bayesian Algorithm</u>

The simple Bayesian system produced a differential accuracy of 91%, for the test cases of BCC and SKE. The full diagnostic flow chart [5.b) (ii)] above produced an accuracy of 78%. The accuracy could be increased, however, by sequentially pruning the branches where prediction relied upon small sub-sets. The maximal accuracy produced was identical to that produced by the Bayesian method.

c) Taking Associations Between Variables into Account Inference by Substitution of Combined Frequency Estimates

A method of dynamically substituting combined frequency estimates during Bayesian prediction of diagnosis has been compared with a simple Bayesian model for cases of 'suspected appendicitis' and a mix of BCC/ SKE cases.
In suspected appendicitis a 'one out' test method was used. Training and test sets were used for the dermatology cases.

(i) Pair Substitution

The 'pair' substitution method proved more accurate than a simple Bayes model for both sets of cases. In 'suspected appendicitis' dynamic substitution improved accuracy by 9% from 74% to 83%. In the differentiation of basal cell carcinoma from solar keratosis, dynamic substitution increased accuracy from 91% to 95%

(ii) Triplet Substitution

The test was repeated with dynamic substitution of 'triplets'. The accuracy rate of the model fell to 73% for the 'suspected appendicitis, cases and to 91% for the dermatology cases.

d) Production of a DERMIS Reduced Dataset by Elimination of Redundant Variables

The 'reduced dataset' produced for DERMIS by elimination of redundant variables was formed into a single page data collection sheet [Figure 14].

e) Determination of the Number of End-Points to be Used in Prediction by DERMIS

Within the database the number of cases allocated to each of 182 diagnoses groups varied considerably from the hundreds of 'eczema' cases to rare diagnoses containing only one or two cases. The overall diagnostic accuracy, predicted by the Bayesian algorithm when all 182 diseases were considered independently was found to be 55%.

Crossover Between Groups

Analysis of the computer's errors revealed the existence of disease sets. Within each of these sets misdiagnosis (crossover) occurred more frequently than between sets. For example, crossover frequently occurred between the members of the 'eczema like' diseases [Figure 15].

FIGURE 14 Reduced Dermatology Data Collection Sheet

DERMIS: Patient History and Examination Details

Sex

male
female

Age

a5
a10
a20
a30
a40
a50
a60
a70
a+70

Episode Duration

-week
+week

Illness Duration

d1
d7 (days)
d14

m1
m2 (months)
m6

y1
y2
y5 (years)
y10
y+10

continuous

Skin Type

hydration normal
greasy skin
dry skin

Specific History

-psoriasis
+psoriasis

Associated Factors

+pigment change
-pigment change

-itch
mild itch
moderate itch
severe itch

Number Lesions

single lesion
multiple
rash
-lesion

Symmetry

+symmetry
-symmetry

Change Size

increase size
same size

Size (mm)

-size
s10
s20
s30
s40
s50
s+50

Areas Involved

neck
mouth
face
ear
trunk
genitalia
arms
hands
legs
feet
eye lids
nails
scalp
palm/sole

Types Found

macule
patch
papule
nodule
plaque
vesicle
pustule
weal
ulcer
scar
comedone

Colours

-colour
pink
red
purple-pink
white
cream
orange
yellow
gold
light brown
dark brown
black
+colour

Shape

round
oval
annular
pedunculated
irregular
-shape

Border

definite
variable
indistinct
raised edge
active edge

Surface

-surface
warty
scaly
exudate
crust
friable
atrophy
flat & shiny
streaks of white
lichenified
excoriated
umbilicated
+surface

Vascular

erythema
purpura
telangiectasia
-vascular

Palpation

-palpation
soft
firm
hard
rough

Scratch

-scratch
+scratch
++scratch

NOTES : DERMIS diagnostic advice system (keyboard)

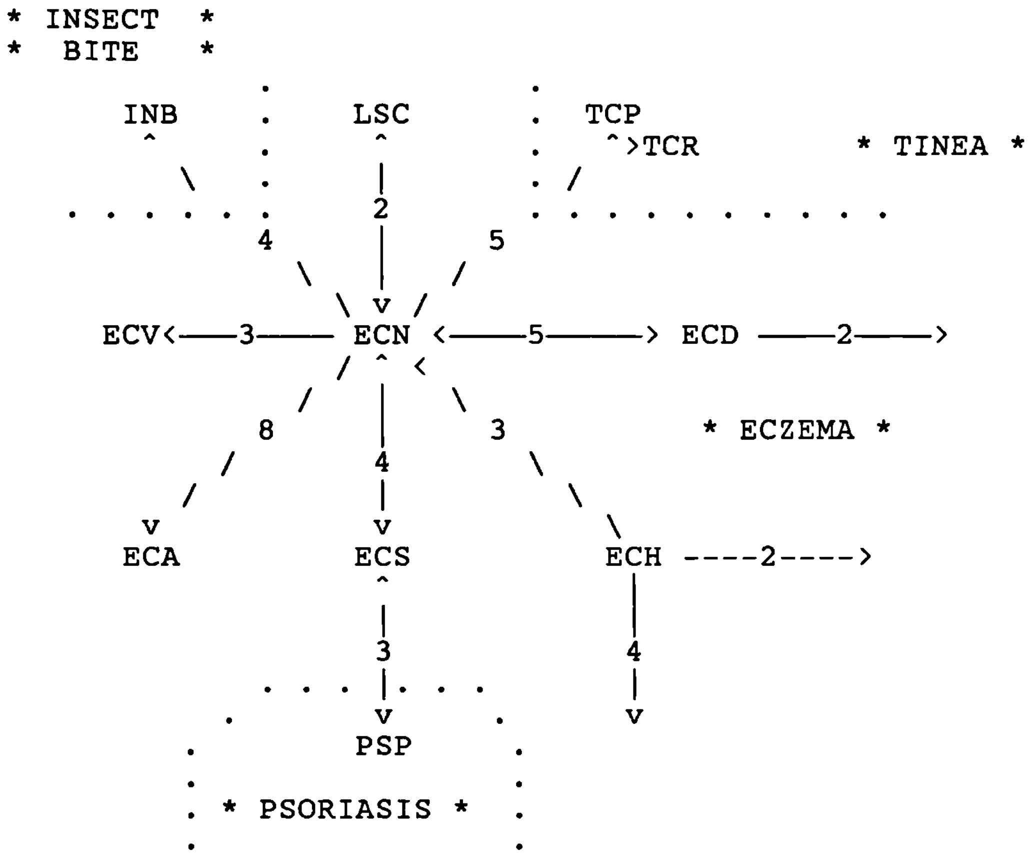
To start program type 'dermis', F1 selects advice routine. F2 to return to summary.

Type codes from data sheet in any order. Follow by 'ENTER'. It will recognise the first part of each code. Repeat code to delete. Type heading for help. Short differential strongest prediction. Try (on this version) F3 compare diseases and F4 game (difficult)

GJ Brooks(10/90)
(0705 584616)
RE Ashton

Figure 15

Schematic Representation of Crossover Effecting the Eczema Set



Where;

LSC	=	Lichen Simplex Chronicus	TCP	=	Tinea Corporis
TCR	=	Tinea Cruris	ECD	=	Discoid Eczema
ECN	=	Non-Specific Eczema	ECH	=	Hand Eczema
ECS	=	Seborrhoeic Eczema	PSP	=	Psoriasis
ECV	=	Varicose Eczema	ECA	=	Atopic Eczema
INB	=	Insect Bite			

and the numbered arrows represent mis-diagnosed cases.

The individual disease groups were formed into end-point groups. In assessing the composition of major predictive end-point groups, such as 'eczema', consideration was given both to the 'between group' failure rates and the clinical acceptability of combination. The result of this review was the production of a 32 end-point group model. [Table 21] gives examples of some of the group combinations involved.

Table 22

Production of the 32 End-Point Group Model for DERMIS: Examples of Diseases Included in Groups

32 Disease Model End-Point Group Name	End-Point Group Also Contain Cases of;
lopecia Areata	Alopecia Totalis
cne Vulgaris	Acne Excoriee, Perioral Dermatitis
czema, Non-Specific	Eczemas; atopic, contact allergic, discoid foot and hand, intertriginous, impetiginised, pomphylx, craquilee, papular , varicose, acute, seborrhoeic, contact irritant
	Lichen Simplex Chronicus, Juvenile Plantar Dermatosis
nsect Bite	Papular Urticaria
ichen Planus	Lichen Planus Hypertrophic
alignant Melanoma	Superficial and Nodular
aevus	Junctional, Blue, Compound, Halo, Hairy Pigmented, Intradermal, Linear Epidermal, Pigmented, Warty Epidermal
soriasis	Plaque, Guttate, Intertriginous, Nail, Scalp
olar Keratosis	Cutaneous Horn
inea	Corporis, Cruris, Incognito, Manuum, Pedis, Unguum
art	Viral, Filiform, Genital, Plane
erruca	Corn, Exostosis

When predictive accuracy tests were repeated with the criterion for success being allocation to the correct one of 32 clinically selected end-pont groups after application of the algorithm, the overall accuracy increased to 68%. In a further test, the 32 end-point groups were formed prior to use of the predictive algorithm. The diagnostic accuracy was again found to be 68%, but the pattern of success and failure within the 383 test cases varied between the two methods.

Details of the number of cases in the database and the test set by diagnostic grouping are shown in [Table 23]. In the table, the name of the test group is followed by its three letter computer code. The 'group size' column gives the number of cases of the end-point group in the database, whereas, the 'test no.' column gives the number of fresh cases tested on the system. The number of cases of each end-point group correctly diagnosed by the system is recorded in the 'no. corr' column. In the final column, lists are given of the cases in each test group that were incorrectly diagnosed by the system. Three letter codes are used to indicate the predicted diagnosis for each case.

Table 23: 32 Clinical End-Point Group Model: Disease categories and accuracy; Groups Formed Prior to Diagnostic Prediction

Diagnostic group	Code	Group size	Test no.	No. corr	Computer prediction if diagnosis wrong
Alopecia areata	(AAR)	23	2	2	
Acne	(ACV)	133	21	21	
Basal cell carcinoma	(BCC)	143	28	27	NSP
Seborrhoeic wart	(BCP)	112	3	2	NCP
Solar keratosis	(SKE)	139	19	15	NID 2SCC REM
Epidermoid cyst	(CYE)	25	1	1	
Dermatofibroma	(DFM)	25	5	5	
Eczemas	(ECN)	371	59	46	ECE 2LSC 4TCP 6REM
Hand & foot eczema	(ECE)	139	19	16	2TCP REM
Granuloma annulare	(GAN)	25	0		
Lentigo	(LEN)	15	4	4	
Lichen Planus	(LPL)	31	4	4	
Lichen simplex	(LSC)	28	3	1	ECN ECE
Molluscum contagiosum	(MCN)	25	5	3	CYE WTV
Malignant melanoma	(MMN)	19	0		
Compound naevus	(NCP)	56	5	2	2NID NJN
Intradermal naevus	(NID)	50	13	5	2BCP CYE DFM MCN 3NCP
Junctional naevus	(NJN)	19	7	5	MMN NCP
Spider naevus	(NSP)	23	1	1	
Pyogenic granuloma	(PGR)	20	4	3	SCC
Psoriasis					
plaque	(PSP)	242	39	31	AAR 6ECA LSC
hand & foot	(PSE)	8	0		
pustular	(PPP)	19	1	1	
Pityriasis versicolor	(PVR)	33	7	5	NJN REM
Rosacea	(ROS)	24	4	3	REM
Squam's cell carcinom	(SCC)	15	2	1	SKE
Skin tags	(STG)	19	2	0	2NID
Tinea	(TCP)	31	8	8	
Urticaria	(URT)	38	3	3	
Warts	(WTV)	122	32	23	4BCP 2NID NJN STG REM
Verruca	(WTS)	29	2	2	
Remainder	(REM)	537	80	20	ACE 5BCC 2BCP SKE CYE 4ECE 24ECN 4LPL 8LSC NID 2NJN NSP PGR PSG PVR 4SCC STG 2TCP WTV WTS
Total		2538	383	260	
Overall Diagnostic Accuracy = 260/383 = 68%					

A review of the patterns of success and failure between the 182 end-point groups and between the 32 end-point groups selected on clinical grounds lead to a revision in major group classification and production of a 42 end-point group model.

f) Selection of a Lower Frequency Bound Estimator

Three methods of estimating the lower bound were tested in sequence and the effects upon system diagnostic accuracy observed for the 182 disease group model. The baseline method used a estimated lower bound probability of 0.1 when no information was available concerning the relationship between a feature and a disease. When this estimator was applied in database formation the accuracy of Bayesian prediction was 55% [4.5.c].

Use of a second estimator of 1/2n (n=number of disease group cases) lead to an increase in accuracy of 1%, but use of the Perks estimator increased accuracy by a further 2% to 59%.

g) The Representation and Reliability of Expert Beliefs

The accuracy of classification of a set of clinical algorithms designed to assist in the diagnosis of three common diseases, psoriasis, solar keratosis and basal cell carcinoma has been compared with that of the DERMIS program on a dermatology database of 5336 prospectively collected case records.

Of the 5336 records, 446 were from cases of psoriasis, 265 solar keratosis and 319 records from patients who had suffered with a basal cell carcinoma.

The overall accuracy of the diagnostic algorithms and of DERMIS are shown in [Table 24].

Table 24
Comparison of the Overall Diagnostic Accuracy of Clinical Algorithms for Three Skin Diseases, and the DERMIS Program on 5336 Case Records

		Algorithms Prediction		DERMIS Prediction	
		disease	not disease	disease	not disease
Test Case	disease	402	682	845	185
	not disease	405	3901	189	4117

The algorithms detected 39% of cases of the three diseases which they were designed to identify. The DERMIS program made the correct decision on 82% of occasions. On 22 occasions, the algorithms correctly identified a case of one of the three diseases which the program had missed, whereas the program correctly identified 465 cases that the algorithms had missed.

The algorithm for detection of basal cell carcinoma missed more than 50% of cases of the disease. The DERMIS program included the correct diagnosis in the top three of its differential for 98% of the basal cell carcinoma cases.

6) Laboratory Tests of the Performance of the DERMIS System

A series of tests of diagnostic prediction have been made by 'one out' analysis using the dermatology database of 5203 cases and various configurations of DERMIS. Comparisons have been made by overall accuracy using the disease assigned the highest relative likelihood as the system prediction.

The results of comparison with an initial configuration of the system that comprised 221 disease groups with a full data set and lower frequency bounds set by Perks's estimator are as follows;

a) 221 Disease Groups, the Reduced Data Set, Perks Estimator

Overall accuracy = 60%

b) 42 Clinical End-Point Groups, Reduced Data Set, Perks Estimator

Overall accuracy = 72%

c) The Inclusion of Combined Frequency Estimates

The model as in [5.6.b] above, with the addition of fixed combined frequency replacement, where appropriate;

Overall accuracy = 76%

On 95% of occasions the correct diagnosis appeared in the top three of the differential list. A breakdown of performance accuracy by end-point is given in [Table 25]

Table 25

42 Clinical End-Point Group DERMIS: Accuracy of Assignment of Correct Diagnosis to Top of Differential List in 'One Out' Analysis of 5203 Cases in Database

Group Title	No. Cases Database	No. (%) of Times Top Differential	Note
Alopecia areata	40	39 (98)	
Acne	253	232 (92)	
Basal Cell Carcinoma	288	201 (70)	1
Superficial BCC	24	23 (96)	
Bowen's Disease	27	22 (81)	2
Chondrodermatitis	31	28 (90)	
Cyst Epidermoid	59	51 (86)	
Dermatofibroma	90	77 (86)	
Eczema	1124	847 (75)	
Folliculitis	23	19 (83)	
Granuloma annulare	47	42 (90)	
Herpes simplex	23	22 (96)	
Insect bite	31	24 (77)	
Keratoacanthoma	25	22 (88)	
Lentigo	38	28 (74)	
Lichen planus	64	54 (84)	
Malig. melanoma nod.	21	18 (86)	3
Malig. melanoma sup.	51	49 (96)	4
Milia	23	20 (87)	
Molluscum Contagiosum	47	45 (96)	
Naevus	396	250 (63)	
Naevus Spider	36	36 (100)	
Pyogenic Granuloma	47	46 (98)	
Pityriasis Rosea	19	18 (95)	
Psoriasis palm/plant.	10	10 (100)	
Psoriasis	414	379 (91)	
Pityriasis Versic.	68	54 (79)	
Rosacea	60	53 (88)	
Scabies	52	51 (98)	
Seborrhoeic Wart	266	183 (69)	
Solar keratosis	269	187 (69)	
Skin Tags	33	32 (97)	
Squamous Cell Carc.	35	26 (74)	2
Tinea	82	64 (78)	
Urticaria	69	65 (94)	
Vitiligo	24	23 (96)	
Warts	277	222 (80)	
Verruca	50	49 (98)	
Other Single Lesion	217	109 (50)	
Other Multiple Lesn.	190	101 (53)	
Other Rash	253	104 (41)	
No Rash or Lesion	7	6 (86)	

Notes on Accuracy Figures in Table 25

1. 49 failures were diagnosed as other malignancy
2. 4 failures were diagnosed as other malignancy
3. 1 failure was diagnosed as a superficial melanoma
4. 1 failure was diagnosed as other malignancy

The accuracy with which the diagnosis was predicted varied between groups. For example, 847 (75%) of the 1124 cases of eczema were correctly identified by the program, compared with 49 (96%) of the 51 cases of superficial spreading melanoma. Cases of rarer disease, e.g. pemphigoid, mycosis fungoides, which were assigned to 'remainder' groups, made up 13% of the total database. These groups had the highest failure rates, i.e. 41-53% (Table 26).

Table 26
Most Common Errors: Confusion Between Specific End-Point Groups

Actual Diagnosis	Top of Differential	Number of Cases (%)
Eczema	Psoriasis	84
Send to Specialist	Eczema	70
Eczema	Scabies	63
Eczema	Tinea	38
Naevus	Seborrhoeic Wart	21
Eczema	Pityriasis Rosea	20
Eczema	Send to Specialist	20
Naevus	Superficial Melanoma	18
Basal Cell Carc.	Squamous Cell Carcinoma	18
Naevus	Skin Tag	18
Send to Specialist	Pityriasis Versicolour	17
Seborrhoeic Wart	Naevus	15
Acne	Rosacea	15

The most common cross-overs between groups are given in [Table 26]. The eczema group produced the most failures amongst the groups that have been assigned disease names. There was cross over between eczema and a set of diseases that can have a similar appearance; psoriasis, scabies, tinea and pityriasis rosea.

Amongst the malignant tumours, cross over was again commonly to similar looking diseases. For example 49 basal cell carcinomas were mis-diagnosed as other malignant tumours.

d) Application of Expert Belief to Lower Bound Estimates

The model as in [4.6.c] above, with the addition of expert beliefs in lower bound determination;

Overall accuracy = 83%

On 97% of occasions the correct diagnosis appeared in the top three of the differential list. During testing the expert belief 'rules' regularly excluded 70% or more of end-point groups from appearing in the differential.

7) Trials of the DERMIS System in Clinical Practice

Semi-field testing of the DERMIS system has been carried out in order to investigate the potential problems and implications of implementing the system in primary care. The following issues have been addressed;

a) The Advice Required by General Practitioners Compared with the Advice Available From DERMIS

A further survey of the referral habits of general practitioners has been conducted amongst 125 cases randomly selected from the dermatology database. In each case appropriate information has been extracted from general practitioner's referral letter. Summary tables been compiled where the reason for referral has been compared with management outcome and DERMIS advice (based on 'one out' calculation).

A sample of 125 randomly selected dermatology case records was scrutinized in detail, in order to determine the reason for referral and the outcome of specialist review. The results appear in [Table 27].

Table 27

A Random Sample of 125 Cases Referred to Dermatology Clinics By Primary Care Physicians: Reason for Referral vs. Outcome

Reason for Referral	Outcome			
	Malignant Tumour	Benign Lesion(s)	Rash No Infection	Infection/ Infestation
Diagnosis & Management				
... diagnosis unknown	* 5	* 19	* 21	* 7
... ? malignant	* 6	* 11	* 3	-
... ? infection	-	-	* 3	-
2nd Opinion (correct) (diagnosis)	-	* 4	* 2	* 1
Management	-	4	2	3
Further Treatment	-	-	15	9
Removal/Biopsy	-	10	-	-

* Indicates the cases shown in [Table 28]

In 76 (61%) of cases [Table 28] the general practitioner requested assistance with diagnosis. The final diagnosis matched one of the 38 main DERMIS end-point groups in 71 (93%) of these cases. DERMIS (using 'one out' analysis) placed the correct diagnosis at the top of its differential list on 54 (71%) of the occasions that general practitioners had requested diagnostic assistance.

Table 28
Breakdown by Diagnosis of 76 Randomly Selected Cases Referred by General Practitioners to Dermatology Clinics for Diagnosis and Management

Final Diagnosis	Number of Cases
Basal Cell Carcinoma	9
Bowen's Disease	1
Discoid Lupus	1
Eczema	15
Erysipelas	1
Insect Bites	2
Lentigo	2
Lichen Planus	2
Naevus	12
Rare Tumours	2
Pyogenic Granuloma	1
Pityriasis Rosea	1
Psoriasis	4
Pityriasis Versicolour	1
Rosacea	1
Scabies	2
Solar Keratosis	7
Squamous Cell Carcinoma	1
Tinea	5
Urticaria	2
Vasculitis	1
Viral Wart	1
Verruca	2

b) The Choice of User Interface and Explanation Routines

Three methods of inputting data to DERMIS have been tested in day to day use in the dermatology clinic. The results are as follows;

(i) Doctor completes a data collection sheet then enters numerical or mnemonic codes into the computer, via the keyboard.

- This method was abandoned as it was found to be too slow for real time use.

(ii) Doctor completes a data collection sheet that rests on a touch sensitive input device. Ticks are sensed by the device and codes automatically entered into the computer.

- This was the fastest method of data entry tested. Standard cases could be entered in less than a minute by inexperienced operators. Doctors and students favoured pen input, but the equipment was too unwieldy, and lacked sufficient portability,for routine use. The method may be used in future.

(iii) Doctor uses a keyboard to select appropriate answers from menus using single key presses.

- Keyboard entry was found, by users to be satisfactory if the number of key presses required to operate the system was kept to a minimum. A trained user can enter the details of a case in less than a minute using the current keyboard entry system.

Of various explanation routines that have been tested by students and trainees;

(iv) The user is able to add data to, and subtract data from, the case record in any order. He is presented with an immediate update on the relative likelihood output.

- This method allows rapid hypothesis testing, with the user in control of the process. If no order to data entry is assumed then doctors can adopt their usual sequence of examination. This method of operation has been popular with all users.

(v) Production of ordered lists of features that support any selected diagnosis rather than any other, or rather than all the rest. The method employed was a development of that described at [3.2.f].

- This method has been abandoned as it was not used in practice.

(vi) Production of a ranked list of features that are critical to the order of the current system differential. This is produced by an iterative process, conducted for each feature entered, of recalculating the differential, with the postulate that the feature has been removed. The result is produced in less than 5 seconds on an IBM compatible 286 portable running MUMPS.

- This is a new method that has been welcomed by the dermatologist. No user feedback is currently available.

(vii) Production of a list of diseases excluded by expert opinion and features that have caused exclusion.

- The dermatologist uses this routine regularly to test that the beliefs represented in the system match clinical presentations.

The dermatologist has reviewed the differential diagnosis produced by the 42 end-point model of DERMIS for 50 sequential clinic cases at the time of presentation. His opinion was that the conclusions reached were all reasonable reflections of the clinical material.

c) Semi-Field Trial of Dermis as a Decision Support Tool for Primary Care

In a semi-field trial of DERMIS, 49 general practitioners, and 9 hospital doctors (hereafter referred to collectively as general practitioners) used check lists when collecting clinical information about 8 dermatology patients. For each case, the items of information collected by each general practitioner were compared with a list of features collected by a consultant dermatologist ('approved' features). The results are shown in [Tables 29, 30]

Multiple choice data collection sheets were used by the general practitioners. This meant that when an 'approved' feature was not identified by a general practitioner, he would in fact supply alternative information. For example, if the 'approved' colour of a lesion was 'red', the general practitioner might actually supply 'light brown' and 'pink'. Considerable variation has been found, between the cases, in the numbers of 'approved' features collected by general practitioners [Table 29].

(i) Data Collection by General Practitioners

Table 29

Observer Variation: Diagnosis, Number of Data Items Collected and Computer Diagnosis

ase	Final Diagnosis	No. GPs	Cons. Items	GP Average (S.D.)	Computer Diagnosis (Cons. Data)
1	Naevus	46	6	5.0 (1.0)	Naevus
2	Dermatofibroma	57	6	5.1 (0.9)	Dermatofibroma
3	Naevus	35	5	4.3 (0.7)	Naevus
4	Malig. melan.	53	7	4.6 (1.1)	Malig. melan.
5	Squam. carcinoma	52	8	2.7 (0.9)	Squam. carcinoma
6	Psoriasis	36	7	4.1 (0.8)	Psoriasis
7	Insect bites	43	5	3.0 (1.2)	Insect bites
8	Tinea cruris	40	6	3.9 (1.0)	Eczema

where the columns denote;

Case - contains the trial index number of each case

No. GPs - gives the number of general practitioners observing each case

Cons. Items - gives the total number of features collected by the dermatologist in each case

GP Average - gives the average number of features collected by general practitioners for each case, S.D = standard deviation

Computer Diagnosis - gives the diagnosis of the computer using the dermatologists data from each patient.

There was also apparent variation, between the cases, in general practitioners' ability to identify the same clinical feature. For example the clinical feature 'round' was 'approved' for both case 2 and case 5. Whereas 96% of the general practitioners who examined case 2 observed the lesion to be round, only 23% of those observing case 5 made a similar observation [Table 30].

The fact that a particular feature was observed in a case was more important on some occasions than others. In [Table 30], the cells marked with a '*'indicate instances where identification of a feature was critical to the ordering of the computer's (top three) differential diagnostic listing. In case 5, for example, failure to identify the feature 'crust' was likely to affect the advice produced by the computer. In case 6, failure to observe the (incidental finding) of crust did not affect the ordering of the computer's differential diagnosis.

Table 30
Observer Variation: Comparison of the Ability of General Practitioners to Identify the Same Features of Diseases as a Dermatologist.

'Approved' Features	Patients (cells contain % of observers identifying)								
	1	2	3	4	5	6	7	8	All
Round		96*	51		23		90		64
Pink	78*	54			58	54	63	46	58
Papule	98*	96*	92*				78*		92
Crust					81*	0		80	58
Scale		88*		89		95			90
Warty	89*	84							87
Defined border	98*	96*	86	84	23*	59*	49	56	70
Dark brown			95*	80*	0				54

where: Patient columns 1 to 8; give % of observers of each case who identified the indicated features as being present

Features column; gives a selected list of features found by the dermatologist when examining the patients.

A blank cell indicates that a particular feature was not found to be present by the dermatologist in the case

All column; gives the % of occasions, for all cases, that the dermatologist's observation of a feature was repeated by general practitioner observers. '*' after a percentage indicates that the feature was of particular importance in determining the position of the marked disease within the (top three) differential diagnosis listing produced by the computer for the case.

The decisions concerning diagnosis and management of the 8 cases made before receipt of DERMIS advice have been compared with those made after advice had been given. Comparisons have also been made with the 'true' diagnosis and the dermatologist's recommended management. The results are shown in [Table 31].

Where:

disease names have been abbreviated as follows;

Seb. wart	= Seborrhoeic wart
Dermatofibr.	= Dermatofibroma
Malig. melan.	= Malignant melanoma
Squam. carc.	= Squamous cell carcinoma

Table 31: Effects of Use of the DERMIS System Upon General Practitioner Diagnostic and Management Decision Making

'ase	Final Diagnosis	Unaided GP Diagnosis	Computer Acc % GP data	GP Acc% Before Advice	GP Acc% After Advice	Change Approp Manage	Change Total Refer
1	Naevus	Seb. wart Naevus	61	30	65˜	+5	-5
2	Dermatofib.	Dermatofib. Melanoma	89	46	70˜	+2	-2
3	Naevus	Naevus Seb. wart	97	94	100	0	0
4	Malig. mel.	Malig. mel. 'various'	92	85	94∞	+1	+1
5	Squam. carc.	Squam. carc. Psoriasis	83	58	90˜	+7	+7
6	Psoriasis	Psoriasis Tinea	89	47	67■	+7	0
7	Insect bites	Insect bites Folliculitis	56	30	67˜	+11	-2
8	Tinea cruris	Herpes Impetigo	25	0	27˜	+8	0

the columns;

Unaided GP diagnosis - gives the two diseases most frequently identified by general practitioners as being present, 'various' indicates that a large variety of opinions were expressed

Computer Acc % GP data - gives the % accuracy of the computer diagnosis using data collected by the general practitioners

GP Acc% Before Advice - gives the % diagnostic accuracy of general practitioners before receiving computer advice

GP Acc% After Advice - gives the % diagnostic accuracy of general practitioners after receiving computer advice,
McNemar's test applied, ˜= p<.001 , ■ =p<.01 , ∞=p<.05

Change Approp Manage - gives the number of cases where the management recommendations changed from being inappropriate to appropriate following computer advice.
Change Total Refer - gives the difference between the total number of referrals recommended before and after receipt of computer advice. e.g -5 indicates 5 fewer referrals recommended after advice

For each case studied, the proportion of general practitioners who made the correct diagnosis increased following receipt of computer advice. In 5 of the cases, the improvement was highly significant ($p<.001$). Only one example could be found of a general practitioner changing from a correct to an incorrect diagnosis following receipt of the computer's advice.

In scoring the changes in diagnostic ability, no account has been taken of general practitioners who made an incorrect initial diagnosis, realised this following computer use, but were then unable to decide what the diagnosis should be.

It proved difficult to produce 'Gold standard' management plans for some of the test cases as, for example, the decision to remove a benign skin tumour might be made on cosmetic rather than clinical grounds. Some general practitioners changed their clinical management plans following receipt of computer advice [Table 31]. In these cases, saved referrals of benign tumours have been considered to demonstrate improved management.

In all but one patient, the proportion of general practitioners producing appropriate management plans increased following computer advice.

The referrals planned by general practitioners have been considered separately [Table 31]. The was very little difference in the total number of referral recommendations made before and after use of the computer. Where malignancy was identified more referrals would occur. Where malignancy was ruled out fewer would occur. Where making the appropriate diagnosis lead to a different choice of medication, the referral rate was unaffected.

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Chapter 5

Discussion and Medical Context of Results

Experimental Work Performed; Requirement, Nature and Extent

1) Comparison of Inference Models for Acute Chest Pain Diagnosis

The purpose of the investigation has been to carry out an independent comparison of several established acute chest pain diagnostic advice systems in order gather information concerning the relative performance and applicability of different inference models used in the same clinical setting.

A Comparison of the Chest Pain Advice Systems: Discussion of Results

As we have seen, ACP can indicate the onset of a potentially fatal medical condition. Casualty officers have to be able to decide amongst other things;

- how ill is the patient?
- should the patient be admitted?
- should any treatment be given? eg thrombolytics, analgesia
- should the patient be admitted to CCU?

As part of the experimental work, several advice systems that have been designed to assist the casualty officer with these decision making tasks [Table 3] have been simultaneously evaluated, using a standardized test set of cases.

Before any comments can be made about the findings, several issues, concerning the methodology must be addressed. A major problem with ACP advice system construction and evaluation has been the definition of final diagnosis. In patients who are discharged from the casualty department, the final diagnosis may never be known. In other cases, the extent of investigation, will depend upon local policy and might vary between patients with the same condition. These problems can lead to considerable bias in assumed diagnosis, which in turn will have implications for the accuracy of both disease representation and system evaluation.

In this study, all included patients have undergone rigorous investigation where 'gold standard' criteria have been applied in a consistent fashion in order to accurately establish diagnosis [3.1.c]. However, the patients who make up this group are not representative of all patients who might present to a casualty department as the majority of 'obvious' non-cardiac cases will not have been considered for CCU admission. At least 71 of the 104 cases studied were 'high risk' cardiac patients [Table 5].

As a means of minimising the chance that errors and omissions by observers might influence the quality and quantity of data collected, and hence the validity of any comparisons made between systems, each patient was examined by two experienced clinicians using data collection sheets. In practice, it is unlikely that high quality clinical data will always be available to advice systems located in the casualty department.

In analysing the results, several comparisons have been made between the accuracy of diagnostic prediction by ECG interpretation and through advice system use. In this study, ECG tracings were taken on admission to the CCU using a 12 lead machine and interpreted by an experienced cardiologist. Although, these records are likely to have been temporally consistent with any taken in the casualty department, the quality was normally superior. In addition, the standard of opinion expressed on review of the tracings may not be representative of that of casualty officers.

Where appropriate, systems have been assessed for their ability to identify either acute MI or IHD as a cause for ACP. Efficient prediction of either end point might indicate suitability for use in practical decision making. Goldman, has suggested that only patients with acute MI should be assigned to the high risk category and that other cardiac patients, including those with unstable angina, should be treated on intermediate care wards (where these exist)(90). Early identification of patients who have suffered acute MI might also allow intervention with thrombolytic agents. Identification of patients who are not suffering with ischaemic heart pain might help prevent unnecessary CCU admission (54).

Each of the 104 test case has been assigned to one of three groups by diagnosis [Table 5]. Those who suffered an acute MI can be considered to be 'high risk' cardiac patients. Those who suffered angina are 'medium risk' cardiac patients. The 'other' patients did not suffer cardiac pain.

By using the sensitivity and specificity data from [Table 3], it is possible to estimate the likely accuracy of each of the models, to predict the presence of acute MI or IHD in members of the test group. For example, the estimate, thus obtained for MI prediction, for Pozen's system, is 97% for de Dombal's 95% and for Goldman's i=82% , ii=83%.

The results of comparisons of predictive accuracy following application of the models to the test set appear in [Tables 7,8,9,10]. The designer's test results for each can be found in [Table 3].

As casualty officers are expected to use the systems to assist with their decision making, it seems appropriate to compare the unaided accuracy of doctors with that of the advice systems.

The models of both Goldman and de Dombal have been found to be significantly better than admitting casualty officers at detecting which patients, within the test set, had suffered acute MI [Table 9]. However, Pozen's model has been found to be less accurate than the casualty officers [Table 7].

When the performance was compared with regard to the detection of a cardiac cause for pain, the casualty officers attained higher accuracy than both Joswig's and Pozen's logistic regression advice systems [Table 8]. de Dombal's simple Bayesian model again demonstrated a significantly higher assignment accuracy than all three. On this evidence, it is difficult to see how advice systems such as Pozen's and Joswig's, that appear to be less efficient at predicting outcome than their potential users can be expected to improve human decision making.

The Goldman and de Dombal models have been found to be roughly equivalent in their performance [Tables 7,8] (although Goldman has reported that his system is more accurate (91)). If admission to the CCU had been based upon the recommendations of these advice systems then 3, 6 or 8 cases of acute MI would have been missed depending upon whether the Goldman (ii), de Dombal or Goldman (i) systems had been used.

Several factors could account for the differences found. The performance of a system might well be affected by the mix of patients presented to it. The casualty officers will see more patients with ACP than they admit to the CCU. If such patients could be easily identified by a system, then the determined specificity might well be higher than found here. On the other hand, systems that have been designed to detect patients that have suffered acute MI should be able to detect them wherever the test is performed.

Doctors are taught to ask patients suffering with chest pain about the nature, location and radiation of chest pain. It is of interest that Goldman, de Dombal and Joswig incorporate detailed information of this nature into their predictive systems, whereas in Pozen's model we are asked only if pain is the most important symptom. The are numerous other differences between the number and type of variables used. It may be that differences in performance could in part be due to the initial selection of variables.

The number of variables used in final implementation may be a sub-set of those initially thought to be important. Selection might occur on clinical grounds or through the use of statistical techniques such as regression. The use of a small final data set might result in system being more population dependent than one that used a large list of variables, because the best fit for a particular instance had been adopted. This may apply to the Joswig model which was initially tested upon the data set from which the logistic regression coefficients were derived.

The offered prediction, in one of the small data set models, could also be particularly influenced by poor history taking. For example, the presence or absence of sweating or subjective impressions of the severity of the pain could cause dramatic differences in the prediction produced. These factors could well account for the poor general performance of the Pozen model, when compared with its specifications [Tables 3,7,8].

The Goldman flowsheets are also small dataset models. Should they not also be affected by these factors? It is of interest, that when Poretsky (29) carried out an independent test of the Goldman (i) model on 168 patients who had been admitted to hospital with suspected MI. He found that, whereas, the sensitivity of the physicians and the algorithm were equivalent, the specificity of the physicians was 85% and that of the algorithm 62%. Poretsky's trial had been in New York. He compared the type of practice with

that of Boston where Goldman had first used the protocol. He explained the different results on pressure of beds. In New York, shortage of coronary care beds meant that only 58% of patients with ACP were admitted, whereas in Boston the admission rate was between 90 and 100%. He postulated that the New York physicians were able to raise their specificity without loss of sensitivity to meet the circumstances.

Another factor is, however, is of critical importance in determining the predictive behaviour of Goldman's flowcharts. The first question on the chart [Figure 2] asks whether the ECG changes are suggestive of MI. If the answer is 'yes' then MI is predicted. The decision made will reflect the ability of the advising doctor to detect an MI by reading the ECG, rather than any expertise inherent in the design of the chart. Wyatt has described this as circularity (167)

In the trial, ECGs were read by a cardiologist, who was able to predict the presence of acute MI from the ECG more frequently than the admitting casualty officers. In addition the specificity obtained was higher than that of any of the advice systems. Application of the Goldman charts could well have increased sensitivity through identification of further relevant information in cases where the ECG did not suggests MI, at the expense of some loss in specificity. If Goldman's charts had been used by casualty officers it seems unlikely that such high performance rates would have been obtained.

Other case driven models have been used to detect patients that have suffered MI. Hart and Wyatt (143) compared several configurations of a neural network with a simple Bayesian system during laboratory tests and found that the Bayesian system was more accurate. They also found that its output was easier to interpret. Wyatt rejected the possibility of using a Bayesian algorithm at an early stage in the development of ACORN but then found that the subsequently developed expert system did not offer any advantage to casualty officers (25).

One way of avoiding subjective variation in ECG interpretation is allow direct machine reading and analysis of tracings (215). This is being attempted for ACORN (30,25).

Both the Goldman (i & ii) and de Dombal systems appear to have demonstrated classification performance rates that might render them suitable assistants in ACP triage, although there may be favourable bias associated with assessment of the Goldman system. It is likely that de Dombal's Bayesian model would prove to be the more versatile as it does not require expert information in order to produce a prediction.

To be of benefit, an advice system must pass on to the user, the advantages of its accuracy, without incurring the penalties resulting from missed high risk patients. In practice, both the de Dombal and Goldman (i or ii) systems might assist their users to rapidly identify high risk cardiac patients and enable them to supply appropriate treatment at the optimal time. de Dombal's model might also be of value in the classification of low risk patients that do not require admission. The ultimate decision, concerning whom to admit or treat, however, must remain firmly within the control of the advice system user, as no system has demonstrated 100% sensitivity for high risk patient identification.

The conclusions from the study are that advice systems produced for the same clinical setting will not necessarily perform in the same way. Different designers will not necessarily find the same variables to be important even though they have used a case driven methodology. Various factors may affect performance including data set size, method of optimisation and embedded expertise either in variable selection or system operation. The simple Bayes model and a decision tree offered performance that could be used as justification for further field trials. The logistic regression models did not. The performance of the decision tree relied upon the availability of expert ECG interpretation.

2) Hospital Field Trial of The Leeds Acute Abdominal Pain Diagnostic Advice System

The Leeds acute abdominal pain advice system has been under development for over 20 years. There have been many trials of its performance, including multi-centre hospital field trials. It has been shown to benefit both patients and doctors who use it, yet it has not been adopted for general use [2.10.].

All results to date appear to have been, processed by the developers and it has therefore been of value to conduct an independent trial within a hospital where doctors had not heard of the system and had no particular bias for or against it.

Main objectives have been to assess the system's acceptability to users, its suitability for application to the clinical task and the strengths and limitations of adopted trial design method.

A total of 353 patients were included in the various phases of the trial. Despite the logistic difficulties encountered [4.2.b], comparison with the National trial results, and OMGE survey [Table 11], revealed few differences from the expected proportions of diseases within the study population.

Definitions of diagnostic category were available (206), however despite this, assignment of final diagnosis remained a likely source of error. If a laparotomy was not performed, and the patient recovered then an end-point diagnosis of NSAP was often made. It is possible that some of these were cases of appendicitis or other 'surgical' diseases that resolved, in which case, variation in the time for operative intervention between surgeons could alter the final classification produced.

Following operation, there was frequently conflict within the record as to what had been found. A surgeon could record the removal of an inflamed appendix, which was later reported to be normal by a pathologist examining tissue taken from it. For the purposes of standardization, the pathology report was always considered to be correct. However, even such tests are unlikely to achieve perfect sensitivity and specificity. Confusion might arise, for example, from 'periappendicitis' (15), where the appendix is involved with other abdominal inflammatory disease.

In order to reduce the effects of final diagnosis error, cases were followed up, through hospital record surveillance, for at least one year and in a random sample of cases the diagnostic category assignment was checked by an independent physician.

Another potential source of error has occurred through the 90% compliance rate for form filling during the intervention phase. This finding was later explained during feed back sessions. Form filling only became routine for one house officer. For the other house officers involved, it represented a duplication of effort. They all tended to 'forget' to fill forms in when they were busy [4.2.f].

As previously found (5), use of the advice system, whether it be forms or forms and computer, appears to have a beneficial effect upon the diagnostic classification accuracy of doctors. Significant increases in both HOs and SHOs performance [Table 12] of þ15% occurred during the intervention period. Additionally, there was a significant fall in the negative laparotomy rate from 14.6% to 7.6%. However, the other indicators of performance showed only marginal change.

In contrast to the national trial (5), few savings of resources probably occurred during the trial at Haslar. This may be explained in part by the admission policy. Although the medical officers were perhaps in a better position to make an accurate diagnosis, this could have no effect upon the number of patients admitted as was controlled by referring general practitioners.

When the changes in diagnostic classification accuracy were investigated further [Table 13], it was found that the greatest increase occurred in those correctly identified as suffering appendicitis, although NSAP identification also improved. This change in performance might account, for the reduction in the negative laparotomy rate.

Hall (205) has suggested that the increases in diagnostic accuracy of medical staff using the advice system could be entirely explained by adoption of the data collection forms and additional consultant interest and support. This is a point that was not clearly discussed in the Leeds trial reports (8).

The Head of Surgery at Haslar had not considered the feedback of performance information from consultants to house officers to be appropriate and this part of the abdominal pain system was not ever implemented. Hall's hypothesis that the forms themselves might account for the improvement in accuracy is in fact supported by the findings from the Haslar trial, where no significant difference was found between the accuracy rates of doctors who used forms only and doctors who used both the computer and forms [4.2.c].

Judging from from the users comments [4.2.f], the computer may well have had an adverse effect upon house officer performance as obtaining computer advice required a break in normal routine. As with the data collection forms, the extra effort required for computer use was only tolerated whilst house officers were not busy. Only 56% of the data collection sheets completed during the 'computer access' phase of the trial were actually used to obtain computer advice. The justification routines were not used at all. Three months into the planned final phase of the trial, one house officer left and use of forms and computer was abandoned.

The computer's overall accuracy for the cases where it was used was 67%. It is difficult to see how receipt of its advice could favourably influence management decisions that were actually being made by SHOs, who had demonstrated an accuracy of 79% before the computer was introduced [Table 12].

The Trial Methodology

The use of a multi-phase trial methodology, where the effects of changes of practice are compared with baseline performance figures, introduces several sources of bias. In hospital, the medical officers and ward staff taking part often change between phases. There are likely to be differences in diagnostic performance between doctors. A diagnostic system of little worth might appear to show advantage when used by 'keen' and possibly more expert clinicians. An experienced senior house officer might favourably influence the performance of a junior during one phase but might well depart before the next.

Another problem lies with the "Hawthorne" effect where the performance of those studied, eg medical officers, improves when they know that they are being observed. The effect may occur during the intervention phase but be absent when passive baseline data collection is in progress (5,8). Wyatt and Spiegelhalter have suggested low and high profile baseline phases to measure the effect (167).

In order to counter these criticisms, the Leeds AAP system was field tested at multiple centres (5,8). Trials that involve sequential control, test and cross-over phases can be protracted and difficult to interpret. An alternative is a trial methodology that uses simultaneous controls. This method was carefully applied in tests of ACORN where there was randomization for a 'computer use' or 'control group' following data collection (25). These testing methods may well be suitable for narrow domain advice systems operating in the controlled surroundings of one department of a hospital, eg ACORN or the Leeds AAP system, but could be difficult to apply in a primary care setting (14) or where a system such as QMR (199) or OSM (202) was able to offer advice about a wide range of medical conditions.

3) Comparisons of the Performance of The Leeds Acute Abdominal Pain Diagnostic Advice System with Paramedics, Non-Medical Staff and Referring General Practitioners

There has been little investigation of the possibility of using the Leeds advice system in primary care where non-specialists might welcome diagnostic advice when making decisions concerning patients suffering with AAP. In the Haslar abdominal pain trial, for instance, decisions about admissions were made mainly by general practitioners or casualty officers.

A comparison has been made between the assigned diagnosis of referring general practitioners and that of the Leeds AAP system on a test set of 99 cases of 'suspected appendicitis' [Tables 16,20] whose details had been collected by house officers.

The finding of some 26% difference between the accuracy rates of the referring general practitioners and that of the computer might be used to support implementation of the advice program in general practice.

There are, however, several factors that need to be considered. Many diseases that can present with abdominal pain, such as mild gastro-enteritis, are self limiting and rarely necessitate hospital admission. They may be under-represented in the advice system's database. The symptoms of appendicitis develop with time. In the first few hours patients with self limiting non-surgical disease may be difficult to distinguish on clinical grounds from those who are developing appendicitis.

These factors may well mean that general practitioners are having to make decisions about patients who are presenting with patterns of clinical features that are slightly different to those found in patients with the the same diseases in hospital.

From [Table 14], it appears that general practitioners had difficulty identifying patients who were suffering with NSAP. However, it is likely that they subconsciously raised their sensitivity levels in order to avoid missing any cases that required surgery. It may well be that general practitioners have less current experience in dealing with AAP cases and also have a resulting low classification accuracy.

In some hospitals that took part in the multi-centre abdominal pain advice system trial, the Leeds AAP system was used in casualty departments (5,7). Patients may present to casualty rather than to their general practitioner at an early stage of symptom progression. When the advice system was exposed to data collected in casualty departments it was reported as having produced similar diagnostic performance to that obtained through exposure to data collected by house officers (5,7).

In the comparison of the performance of computer and general practitioners made using the Haslar 'suspected appendicitis' cases [Table 16], the computer attained a significantly higher accuracy rate and provided advice that might have helped the doctor on 30 (30%) of occasions. However, it missed 8 cases of appendicitis that the general practitioners had detected.

If all patients with appendicitis are admitted and a large proportion of patients with NSAP are not referred, then the prior probabilities used in the Leeds AAP system will be inappropriate for primary care. If the prior probability weightings used by the system had been changed in favour of NSAP before the performance assessment on 'suspected appendicitis' cases, then the systems' false negative rate for appendicitis would have increased.

It is concluded that there is evidence to support further testing of the Leeds AAP system in primary care. The prior probabilities and output decision threshold boundaries would have to be adjusted to meet the requirement for high sensitivity in the detection of surgical disease. However, such changes might reduce the specificity of the system to a point where its use may not confer advantage.

Implementation might increased general practitioner specificity and reduce the number of patients with NSAP that are referred for operation. Critical factors for success might be general practitioners' reactions to incorrect (false negative) predictions made by the computer, and finding a method of implementation that was appropriate for general practitioners to use when seeing patients.

Significant improvements in general practitioner accuracy might be obtained, without computer use, through the simple measure of issuing general practitioners with AAP data collection sheets. As in the Haslar trail, the effectiveness of this measure is likely to be limited by the time taken to complete the form. The data items could be designed into a referral slip. It is recommended that this be the first test as it would also allow data collection for study of any differences in disease presentation between primary and secondary care.

Use of the Leeds AAP System in the Remote Location

Another group of medical decision makers who might benefit from access to the Leeds AAP system are seagoing paramedics. At sea, a paramedic may be faced with making a decision about the evacuation of a patient with AAP. The scale of the potential problem has been investigated for RN paramedics in [4.3.b,i]. Trials have been performed to investigate the ability of paramedics to reach the correct diagnosis in patients suffering with AAP. The performance has been compared with that of the computer advice system.

In order to eliminate the 'check list' effect paramedics were provided with data that had been collected by a house officer [4.3.b.ii]. The loss of direct contact, was perhaps ameliorated by concise presentation of relevant information, although þ25% complained that favourite symptoms and signs, (mostly of little predictive value), were missing from the summary form.

The exercise was designed as a simulation of the problem of remote medicine, as found aboard warships. The paramedics had time to examine text books but were not able to seek other expert medical advice. Their overall diagnostic accuracy was greater than for general practitioners, who had actually seen the cases [Table 20]. However, the general practitioners had seen the patients at an earlier stage and did not have the benefit using a data collection sheet. Of the 99 medical decisions made [Table 18], the computer would have been of value in 22 cases and given inappropriate advice in a further 6 cases, which included one patient with early appendicitis. However, in this test, the computer's sensitivity was greater as it correctly identified 5 cases of appendicitis that the paramedics had missed.

At sea, patients have immediate access to medical care provided by a paramedic and may present themselves at an earlier stage of disease than patients ashore. A general practitioner seeing a patient, suffering with acute abdominal pain, in a clinic or on a home visit is under strong pressure to make an admission decision at the time. If he does not, then there may well be a requirement for a time consuming follow up visit. Although the advising paramedic at sea, is usually requested to give decisive early advice concerning prognosis, there is often time, in practice, to monitor a patient before a casualty evacuation decision has to be made. As the perforation rate in appendicitis has been found to be 4% per 12 hour period in the seagoing age group (6), and the computer appears to offer advantages in both sensitivity and specificity, a policy of controlled reassessment by the paramedic and advice system, would seem likely to give greater low risk accuracy of performance than immediate decision making by a paramedic acting alone. It is also of interest that in this example the computer might offer advantage over and above that provided through use of data collection sheets.

In this scenario there is a danger that the computer might be treated as an expert and used to make decisions rather than support them. The accuracy of the computer's output may be dependent upon the mix of cases presented to it and the accuracy of data collection.

An argument that is often employed to caution against use of advice systems by paramedics is that such personnel are unlikely to be able to collect sufficient accurate medical information to allow appropriate computer use. Dickson (208) found that medical students and newly qualified doctors produced more errors in clinical examination than more experienced house officers, whereas history taking was equally well performed by both groups.

In a small test of the ability of non-medical (coxswains) and paramedical personnel to collect medical information from patients suffering with AAP [Table 19], the data items collected by paramedics were compared with those collected by house officers seeing the same patients.

Of the 10 ten cases for which information is available a substantial difference in content was only found in one. Investigation revealed that the particular coxswain had lost the advice notes, which described how he should conduct abdominal examination. In a second case, a patient was evacuated from sea on a coxswain's advice and admitted to hospital where a (negative) laparotomy was performed. A computer prediction from the data collected by the coxswain indicated a high likelihood of appendicitis [Table 19]. The coxswains were meticulous in data collection and took up to an hour to complete each form.

An advantage of facilitating accurate data collection by non-medical staff, other than the potential for computer assistance, is the likely benefit that could be derived by being able to seek expert guidance at a distance. Coding the information collected into a series of numbers or computer data would allow case details to be transmitted for analysis at a central unit where expertise was available.

The extremely limited results give some indication that adequate data collection for computer use is possible by non-medical and paramedical staff, providing that appropriate training, documentation and time are available.

4) The Design and Construction of DERMIS: A Primary Care Advice Dermatology Diagnostic Advice System

(a) The Clinical Requirement for a Dermatology Advice System

It has been found from a survey of 211 patient referrals made to a dermatology clinic by general practitioners [4.4.a], that on 68% of occasions specialist advice was sought for both the diagnosis and management of skin disease. If a dermatology diagnostic advice system had been available to general practitioners then this might have reduced the number referred.

It was dedcided that the actual 'saved' referrals (37% in survey) would be those that were amenable to treatment in general practice [Table 21]. A number might also be saved where there is little to chose between primary and secondary care management. Some benign skin lesions, for example, are removed for social rather than medical reasons. It proved difficult predict what proportion of patients with benign lesions would still have been referred had the diagnosis been known [4.4.a]. This has been investigated in a further survey and clinical trail [4.6.a] [5.6.a].

The evidence supports the hypothesis that a dermatology diagnostic advice system might be of value to primary care physicians and that its implementation could result in fewer unnecessary referrals to dermatology clinics.

The dermatology advice system would have to be able to assist with the diagnosis of diseases that are normally referred. Treatment protocols might have to be integrated to increase the chance of the correct management policy being selected.

Domain Definition

The domain specification chosen for DERMIS includes all instances of skin disease that might cause a primary care physician to seek the advice of a dermatology specialist. The criteria used by general practitioners to make this decision are explored later [5.6.].

(b) Collection of a Dermatology Database

It has not always been easy for experts to describe their knowledge [1.9.]. In dermatology, Haberman et al. (128) reported the development of a rule based expert system, to assist dermatologists with decision making, where a large number of diagnoses were defined by experts in terms of disease related weightings for symptoms and signs. Unfortunately, the system was not used because its diagnostic performance was not as good as those it was designed to assist. A particular problem encountered was relative calibration of the experts' weightings. An effort to resolve the problem was made by adjusting weights according to feedback of information derived from test cases. Earlier attempts at producing dermatology systems have been described by Stoecker (209).

In pursuit of the work related to this thesis, an early decision was made that the dermatology advice system to be developed (DERMIS) would base diagnostic inference upon the knowledge of disease acquired through the analysis of prospectively collected clinical information [p 80] rather than expert opinion.

However, it has not proved easy to uncouple expert opinion. Different experts may select different sets of questions to ask about the same diseases [6.1.] and the choice of the initial data set could well be of importance in determining the final performance of the derived advice system. In the study of AAP, standardization was attempted by taking into account, the opinions of more than 200 surgeons when designing the data collection sheet (12) [Figure 4].

For DERMIS, one specialist and two general practitioner trainees created a list of the basic questions to ask a patient suffering with skin disease. I then searched standard text books in order to find other information that might be of value in diagnosis. The final 'full' [figure 7] list was formed into a data collection sheet. Definitions were written and tested [3.4.b]. The initial objective was to collect as much information as practically possible, in a standardized way, so that important features of diseases, that might be unknown or subconsciously recognised by experts, were not missed.

de Dombal found that observer variation in data collection from patients suffering with AAP could be minimised by adopting standard definitions (2,174). In data collection for DERMIS it was decided only to use supervised and committed observers in order to maintain high standards of data collection, involving minimisation of missing data and rigorous application of definitions.

A potential weakness with this method has been that the main observer could have introduced substantial bias into the database either directly or by influencing those he was supervising. In particular, the dermatologist might have subconsciously fitted symptoms and signs to diseases rather than faithfully recording new case details, because he is normally able to recognise the disease before making any effort to describe it. For example, he might be tempted to to record lichen planus as 'classically' violacious when in fact the rash he sees is red, pink or even green!

A large database of clinical information has been prospectively collected from 5203 patients and used in the construction of DERMIS. The definition of disease end-points has also required the use of expert knowledge and has relied to a certain extent upon expert opinion. It proved impracticable to confirm the diagnosis in every patient by analysis of tissue samples. A compromise solution was that skin samples would be taken when the dermatologist was in doubt about diagnosis and that every decision would be checked at follow up. Analysis of the database has indicated that the dermatologist's initial diagnostic accuracy might well exceed 90% [4.4.b.ii].

5) Investigation of Measures that Can be Taken to Improve the Performance of Diagnostic Advice Systems that Use a Simple Bayesian Model

The theoretical disadvantages of using simple Bayesian inference models in medical diagnostic prediction are well known [1.7.d] [1.7.e] [3.5.e] [1.14] and include;

- inappropriate assumption of conditional independence
- inappropriate assumption of a mutually exclusive and exhaustive set of diseases
- difficulty in estimation of lower frequency bounds
- a lack of representation of deep or expert knowledge
- inadequate justification of results [5.7.b].

The clinical importance of these criticisms has been investigated using the collected clinical information from the three domains studied.

a) The Relevance of Association Between Variables within the Clinical Domains

Associations between the features of diseases have been found in all the domains studied [4.5.a]. It has been suggested that important feature associations within diseases are largely dependent in nature [1.9.e]. In recent times, this has lead to the use of belief networks to represent the knowledge of disease processes and effects within expert systems (133,134,135,136). Séroussi demonstrated a slight increase in diagnostic performance in an Bayesian system modelled on the Leeds AAP system that had been altered to incorporate the Lancaster model for first order association (50).

Studying the first order associations of symptoms and signs within diseases of the three domains has been rather like looking at a jigsaw puzzle. The pieces all fit together to make a complete entity. Within the whole, groups form meaningful patterns.
Some patterns of feature association have occurred because information about the same concept has been gathered in different ways (59). For example,

[Figure 8] the presence of an 'abdominal scar' and a history of 'abdominal surgery' are both indicators that an operation has actually taken place.

[Figure 11] lesion 'size 1-9mm' and identification of a 'papule' are directly related because the definition of a 'papule' states that the lesion should be less than 1 cm in diameter.

These relationships are largely disease independent and occur wherever the features occur together. However, they are not alternatives. Abdominal scars can be caused without operation and papules come in many sizes (<1 cm) and must be 'raised'.

The effects of using both 'abdominal scar' and 'abdominal operation' as independent variables in a simple Bayesian system might be that;

(i) If the likelihood of previous operation were same for each of the possible causes of AAP then relative positions would be unaltered.

(ii) If, however, sufferers of, for example, small bowel obstruction were more likely to have had a previous abdominal operation than sufferers of NSAP, then the presence of both features would unduly enhance the relative position of small bowel obstruction.

(iii) Alternately, if only one of the features was present, and no validity checks were in force, the remaining feature would help to compensate for the missing data item (59).

Perhaps the information we actually wanted to collect was that there had been an abdominal operation where the abdominal cavity had been opened. This feature can be considered to be present if;

-there is an obvious abdominal scar that is not due to a superficial wound

or

-there is historical evidence that an operation has been performed that involved opening the abdominal cavity.

Confounding factors may cause associations between features. Patients with acute chest pain are often given analgesia soon after admission. Injection of diamorphine alters the symptoms [figure 9]. Patients were given diamorphine because they had 'severe pain' in the 'upper half' of the chest. The pain is now 'getting better' and there is 'no nausea' because an anti-emetic was also given.

Another type of association between features occurs when intermediate disease states are present (65,83) or diseases can present in more than one way.

In appendicitis the symptoms and signs can change as peritonism develops [figure 8];

Initially the pain can be 'central' and the patient is 'pale'. As the peritoneum becomes involved the pain localises to the 'right lower quadrant '(RLQ) and becomes 'steady' and is 'aggravated by movement'. The patient becomes 'flushed'.

The dermatologist's description of a classical case of basal cell carcinoma includes;

- a patient aged over 60, who has a single lesion somewhere on the face. The lesion has a raised edge with surface crust and a size between 10-19 mm.

Just over 10% of cases of basal cell carcinoma in the dermatology database presented in this way. It can be seen from [figure 11] that these features have been picked out as being associated. The features of a second common presentation have also been picked out. Small basal cell carcinomas tend to have a 'normal surface' and a 'raised edge'. As they grow, they are noticed as is the 'size change'. The surface breaks down and an ulcer with crust is formed.

It seems then that within each disease group certain sets of features may be important for diagnosis at certain times. There is evidence of underlying mechanistic dependence within some symptom complexes. In other instances there seem to be small independent clusters of symptoms and signs. Treating all of the variables as being mutually independent will certainly not capture the rich description of disease presentation that is embedded in the case data.

b) Decision Making by Iterative Selection of Variables

Diagnostic flowcharts can be produced by partitioning [4.5.b] (48). They represent the available information in a dependent structure. [Figure 12] is a flow chart that has been derived from the 'suspected appendicitis' cases in the abdominal pain database. The selected groups of features have clinical relevance within the diseases and appear to reflect the rich nature of the evidence. Some features such as 'Rebound' tenderness appear in several branches and appear to be strong independent predictors of disease.

Charts for various combinations of diseases have been produced by this partitioning method [3.5.b]. [Figure 13] gives a chart to help distinguish between basal cell carcinoma and solar keratosis (a common differential problem). The performance of this model has been compared with a simple Bayesian algorithm [4.5.b.iii] in order to assess whether a dependent structure could offer advantages by picking out cases where important diagnostic features combinations occurred. The result was fascinating. The complete dependent model was substantially less accurate than the Bayesian model. Branches of the tree were sequentially pruned and the accuracy reassessed. The chart pruned for optimal accuracy identified exactly the same cases as the Bayesian algorithm.

Simple Bayesian algorithms take all variables into account and frequency estimates are based upon disease totals. During construction of the flow charts, the group sizes gradually get smaller as partitioning progresses and there is less evidence to support the selection of variables for further differentiation. The use of multiple order combinations seemed to compromise accuracy. Perhaps both techniques had been tuned in different ways to recognise the same and most common presentations.

c) Substitution of Combined Frequency Estimates

A method of dynamically substituting combined frequency estimates during simple Bayesian prediction has been described [3.5.c] [4.5.c]. Dynamic substitution takes some account of important associations between variables as they are found to occur in cases. If important combinations are not found then features are treated as being independent. Application of the 'pair' substitution method improved the accuracy of prediction of a simple Bayesian model in tests involving both the 'suspected appendicitis' cases and a mixture of solar keratosis and basal cell carcinoma cases. The advantage was lost when 'triplet' frequency substitutes were used in place of 'pairs'.

An explanation for these findings could be that 'pairs' capture some of the important clinical associations within the disease groups in a general way, whereas the relationships are being overstated when 'triplets' are used.

d) The Reduced DERMIS Dataset

The 'full' dataset [Figure 7] that was used for database collection, proved to be too large for routine use by other than enthusiasts. Apparently redundant variables were found by statistical means [3.5.d]. The findings were discussed with the dermatologist to determine whether there were any known clinical grounds for retention. Those found to be irrelevant to diagnosis on both clinical and statistical grounds were abandoned. The 'reduced' dataset [Figure 14] was formed into a data collection sheet.

e) Determination of End-Points for Prediction to be used by DERMIS:Crossover Between Groups

As the number of cases in the dermatology database grew, so d'd the number of diseases represented. At the time of initial testing 182 diseases had been identified. An assessment of performance was made at this stage by splitting the database into training and test sets and applying a simple Bayesian algorithm [4.5.e]. Analysis of the failures demonstrated that although the diagnoses were clinically mutually exclusive, some diseases had much in common and the Bayesian algorithm was unable to separate them. There were, however, patterns to the failures. Clinically identifiable sets could be found within the 182 disease groups [Figure 15]. For example, numerous types of naevus were represented as separate entities. Some of the distinctions within the sets were not important for the proposed primary care use of the DERMIS system. For example, many of the forms of eczema can be managed in the same way.

A 42 End-point group model was chosen in order to optimize diagnostic performance whilst retaining clinical relevance for the management of skin disease in primary care [4.5.d].

Two methods of applying the end-group combinations were tested. It was considered that combination of group members before application of the algorithm could lead to a dilution effect of the important diagnostic features of the sub-groups (66).

Alternatively, combination of the groups following application of the algorithm might lead to under-diagnosis of the smaller sub-groups. On testing, the outcome of the two types of combination was virtually the same [4.5.e] The actual test cases that succeeded differed, however, which was presumably due to the argued reasons.

The 32 end-group configuration included one 'rare' end-group. This was made up of all of the cases of diseases which could not be included in other end-groups and for which too few cases had been collected to allow independent consideration. Some 21% of all cases collected were assigned to the 'rare' end-group. A similar group appears in the Leeds AAP system under the name of NSAP [1.2.a].

The intrinsic problem with rare disease groups that have been formed by lumping together cases of a disparate nature is that the frequencies generated are not usually typical of any known disease;

For example, taking two rare diseases A and B with equal prevalence

	Frequency in Disease %		
Feature (A+B)=rare	A	B	Combined
Rash	89	1	45
Single lesion	1	79	40
Erythema	90	0	45
Colour brown	0	96	48

The combined disease group undervalues important individual disease characteristics such as in A above, a single brown lesion, and allows representation of feature combinations that do not occur within the group such as a brown rash. For this reason, rare disease groups tend to capture unusual presentations from the other main groups, whilst allowing the larger main groups to capture rare disease cases that have some similarity of presentation.

When the disease categories were re-organised into 42 end-point groups, 4 rare disease groups were formed according to whether they contained examples of; 'single lesions', 'multiple lesions', 'rashes' or 'no lesion or rash present', and were found to reduce the cross-over error rate on testing. The 'rare' end-point groups have fittingly been labelled as 'send to clinic' and make up 13% of the total database.

f) Lower Bound Estimators

In producing the frequency database for DERMIS, from case information, numerous instances were found when no information had been collected that indicated whether a particular feature occurred within a disease. On such occasions a lower bound estimator can be used in the Bayesian calculation. Three estimators have been tested and one suggested by Perks [3.5.f] selected for inclusion in DERMIS following demonstration that its use slightly increased the performance of the prototype version of the program [4.5.f].

g) The Representation and Reliability of Expert Beliefs

Belief networks have been suggested and in some cases implemented as a means of representing and explaining expert knowledge within advice systems [1.9.e]. They have used been as dependent frameworks for probabilistic inference based on the assumption that they can adequately describe the clinical mechanisms involved in their domains of operation (135,136).

Expert opinion has been harnessed at various stages in DERMIS construction [5.4.b]. Following investigation of the occurrence of associated features within the database [4.5.a] [5.5.a] [5.5.c], and the conclusion that a range of distinct presentations might occur within a disease group (particularly the 'send to clinic' group), it was considered that expert beliefs concerning particular presentations might be incorporated into DERMIS and used to tune the inference mechanism.

The beliefs of the dermatologist concerning the relationships between symptoms, signs and particular presentations of diseases had already been published in a book of flow charts designed to assist primary care physicians with diagnosis (44). The charts have been produced independently of work on the dermatology database and use no statistical information derived from case study.

However, the same terms and their definitions have been used. Some interesting 'rules of thumb' appear in the book s ch as, if the nodule appears to be 'stuck on' then it is likely to be a seborrhoeic wart.

For three common disease, psoriasis, solar keratosis and basal cell carcinoma the charts used only terms found in DERMIS [3.5.f]. A comparison of the predictive accuracy of the charts with DERMIS for the three diseases revealed that the DERMIS system was more than twice as accurate [4.5.f]. In addition fewer serious errors, such as missed tumours, had been made.

As the dermatologist's diagnostic accuracy had previously been estimated as exceeding 90% [4.4.b.ii], it has been concluded that the charts do not adequately represent his knowledge of the subject. The detailed relationships described in the flow charts have not been incorporated into DERMIS as they are at best incomplete and unlikely to increase the diagnostic performance of the system.

These findings cast some doubt on the wisdom of using such detailed expert derived knowledge representation charts or belief networks in other advice systems without adequate validation.

6 Laboratory Testing of the DERMIS diagnostic Advice System

Decisions made concerning the basic configuration of the DERMIS advice system have been discussed in [5.5.]. The basic configuration chosen for laboratory test˙ng was a 'reduced' data set, 42 end-group model that used the Perks estimator for lower bound frequency estimation. A series of laboratory tests of various old and new configurations were performed at a point when 5203 cases representing 221 disease groups had accumulated in the database [4.6.]. The purpose was to reassess the choices made about end-point disease groups and to evaluate the use of combined frequency estimates and the incorporation of simple expert opinion upon diagnostic performance.

A 'one out' testing procedure was used throughout [4.6.]. The reason for this was that subsequent clinical trials would involve a system that included all of the cases in its database. If the training database was substantially reduced for the purposes of laboratory testing then the result would be unlikely to represent the 'true' performance of the complete system (74).

a, b) The advantages of combining disease groups into clinical end-point groups were confirmed by a 12% difference in diagnostic accuracy.

c) The Inclusion of Combined Frequency Estimates

A method of dynamically incorporating combined frequency estimates during simple bayesian calculation was investigated at [4.5.c] [5.5.c] and shown to improve discrimination between pairs of diseases. The method required the substitution of a variable number of paired frequency estimates on a case by case basis. A simpler method involving fixed substitution of important combinations of variables occurring within the possible list of answers to individual datasheet questions has been tested using the 42 end-point model [4.6.b]. A 4% improvement in diagnostic accuracy was found when substitution occurred. It is considered that this may have resulted from a reduction in duplication of evidence used during calculation and account being taken of intra-disease variation in presentation [5.6.a].

The 42 End-point groups and their individual accuracy rates are shown in [Table 25]. The accuracy of prediction varies between groups. The 'send to clinic' end-point groups still produced the highest error rates. On 95% of occasions the correct diagnosis occurred in the top three of the differential produced by the system.

The crossover between end-point groups has again been studied [Table 26]. The diseases being confused with one another are no longer members of disease families. For example, the commonest type of failure occurred in 84 cases o eczema where where the system placed psoriasis above eczema in the differential list. Combining the two diseases into the same group is of no immediate value to any system user as the treatments are different.

Apart from accuracy, another measure of system performance is the rate at which the system makes bad errors. In dermatology, one of the most serious mistakes that can be made is to miss a malignant tumour. The penalty of such an e ror is not as great as say missing an MI in a patient with acute chest pain or a perforated duodenal ulcer in a patient with acute abdominal pain, where there may be early fatality. Malignant skin tumours tend to develop over periods of months or years but can certainly have fatal consequences. Early detection can allow surgical removal before there is metastasis. Notes on the detection of various malignancies by DERMIS have been given with [Table 25].

An argument could be made for applying subjective weights to serious end points in order to increase their chance of appearing in the differential (76). However, analysis of computer failures in precise identification of malignant tumours reveals that on the majority of occasions the computer selects a another malignant tumour and on 99% of all occasions the correct malignancy is mentioned in the top three of the differential. For example, the system correctly identified 49 of the 51 cases of superficial spreading melanoma by placing the diagnosis at the top of its differential. In one of the 3 failures an alternative malignancy was identified, and in the other two, malignant melanoma appeared within the top three places of the differential list and 'send to clinic' at the top. The management in all these cases would be the same, excision biopsy or referral to a specialist.

It has been concluded that the addition of arbitrary weighting system for serious diseases would not offer immediate advantage.

d) Application of Expert Beliefs to Lower Bound Estimates

Setting the frequency of occurrence of a feature in a disease to zero, will exclude the disease from the differential list of simple Bayesian system when the feature is identified as being present. In other words, the system excludes the disease because it is known that the feature and disease cannot co-exist. When DERMIS failures of disease identification were reviewed with the dermatologist, it became apparent that on many occasions there were single reasons why particular diseases should not appear in the differential. For example, "acne does not occur on the legs".

The Perks estimator was assigning a likelihood for feature absence in diseases, given the collected data. In the less common diseases, the estimator was having to enter values into empty cells more frequently than for common diseases. Expert opinion has been used to provide definite zero values where it is believed that features and diseases cannot co-exist. The result of this has been further improvement in system performance. First place accuracy increased to 83% [3.6.c] [4.6.d] and correct solutions appeared in the top three of the differential on 97% of occasions. In use tḥe zero frequency settings regularly exclude more than 70% of possible solutions from the differential. It is of interest that the dermatologist has only confirmed þ40% of the empty cell zero values. No zero settings appear in the rare disease end-point groups which has made them more efficient at detecting unusual presentations and cases of disease unknown to the system.

There are potential dangers with the use of fixed rules of this sort. If observers mistakenly identify an excluding factor, then the computer may be forced to produce an erroneous prediction. Such problems can be ameliorated through provision of justification routines that indicate wl ich single collected or absent features most influence the ordering of the differential [4.7.b].

7 Trials of DERMIS in Practice

a) Survey of the Requirement for and Availability of Appropriate Advice

W en, as a prelude to field trials, a random sample of 125 cases from the dermatology database was reviewed it was found that 76 (61%) had been referred for initial diagnosis or second opinion, which confirmed the findings of the initial survey [5.4.] [4.4.]. The system offered a differential listing with the correct diagnosis at the top ın 54 of these 76 cases [Table 28]. It has been considered encouraging that DERMIS would have been able to assist in 43% of these cases referred for diagnostic advice.

b) User Interface and Explanation

It has been seen in the Haslar field trial of the Leeds abdominal pain system [5.2.] that the system failed to be used in over 40% of cases because of the disruption it caused to normal routine. Use of a datasheet appeared to enhance performance but involved duplication of note taking. Entry of data into the computer involved further repetition of the data items.

Several methods of entering data to DERMIS have been investigated [4.6.b]. The fastest and most natural for users involved use of a pen device to tick answers on a data collection sheet. The information was automatically loaded into the computer. This method ,as well as being entirely acceptable to users, offered several other advantages. It meant that;

- the computer screen was available to supply help information
- data could be entered at the time of interview
- no repetition of data recording was required
- on line checks for data inconsistencies could be made
- advice was available when decisions were being made.

Unfortunately, it would be inappropriate to expect all primary care users to purchase such devices in order to allow use of a narrow domain advice system. A second choice that has proved to almost as fast to use allows single keystroke selection of data items from screen lists. The equipment required to support this version already exists in 70% of all UK general practices [1.5.].

It has been considered by many invstigators that users require advice systems to produce suitable, if not extensive, answer justification [1.9.] [1.14]. The most consistently used of the explanatory mechanisms provided w˙th DERMIS has been a simple routine that allows the case cetails to be rapidly altered so that 'what if? ' hypotheses can be tested [4.6.b.iv]. Other more formal print outs of weights of evidence have not proved popular amongst staff in the dermatology clinic or doctors who used the abdominal paın system in Haslar [1.14] [4.2.f].

An explanatory routine that may be essential for primary care use of DERMIS is one that indicates the items of collected information that are vital to the differential output of the system. This would allow users to check that hese items have been correctly identified. There is c rrently insufficient evidence to indicate whether the routine would be used in practice [4.6.b.vi].

Expert Review of DERMIS Performance

P ior to conducting field trials the dermatologist reviewed t e output of DERMIS for 50 consecutive fresh cases and deemed the output differential lists to be reasonable reflections of the clinical material [4.6.b]. This may seem a subjective and almost trivial piece of evidence, but it is ˙mportant that experts are satisfied that a system is producing output that is relevant to the clinic l problem b ing addressed.

c Semi-Field Trial of The DERMIS System

(˙) Purpose of Field Trial

The term semi-field has been used advisedly. Field tests of D RMIS, conducted in the manner of the Leeds abdominal pain trial (5) or according to Wyatt and Spiegelhalter's double bıind controlled 'drug test' methodology (167) have certainly not been attempted [2.2.]. It was considered that a field trial of DERMIS should be conducted to collect information about the following;

- the likely effects of observer variation upon the accuracy of data collection
- the effect of variations in the amount and quality of data collected upon DERMIS system accuracy
- the effect of providing DERMIS advice upon general practitioners' accuracy of diagnosis
- the effects of changes in general practitioner accuracy upon the likely management of patients.
- the likely effect of any changes in the management of patients upon the rate of referral to specialists

(ii) Requirements

In conducting such a trial the following control measures would be required;

- Randomization of patients
- Provision of matched controls
- 'Gold standard' end points for diseases and data collection
- Control of the 'Hawthorne' effect (5)
- Control of the 'checklist' effect (167) [5.2.]

Other important factors;

- Provision of suitable clinical material
- Logistic limitations.

It is considered that the trial methodology adopted has satisfied the information requirements and necessary control measures [3.6.c] [4.6.c]. The other factors have been taken into consideration.

('ii) Control Measures

T e selection of appropriate clinical material was made from a random sample of records that did not form part of the D RMIS database. The aims of selection were to ensure that clinical photographs matched case descriptions and that a variety of clinical problems were included. No reference was m de to the computer during patient selection in order to prevent favourable bias.

During the trial general practitioners acted as their own controls. The details and photographs of the same fully worked up patients were viewed simultaneously by groups of d ctors to prevent any variation in presentation of the clinical material. The original findings of the dermatologist who saw the case were used as the gold standard for data collection.

Each doctor knew that his performance was being studied, albeit anonymously. Although this may have altered attained a curacy, the effect applied equally throughout the trial. T e advantages of using a 'checklist' were isolated from the effects of providing computer advice by measuring general practitioner performance after collection of da a and again after computer advice.

(v) Discussion of Results

T e dermatologist had observed the presence of between 5 and 8 clinical features for each of the 8 cases studied. The general practitioners showed case dependent variation in the f quency with which they matched the dermatologists ndings. For example in case 5, where the dermatologist had collected 8 items of information, the general practitioners co lected on average 2.7 items that matched. In case 1 the dermatologist collected 6 items and the general practitioners averaged 5 matches [Table 29].

There was also variation in the frequency with which general practitioners identified the same 'approved' feature in different cases. For example, the finding 'round' was detected by the dermatologist as being present in case 2 and in case 5. It was identified by 96% of observers of case 2 but only 23% of observers of case 5.

It appeared that general practitioners found it easier to consistently identify some findings, for example, 'papule', than others such as 'defined border' [Table 30].

On occasions when general practitioners failed to identify one of the dermatologist's 'approved' findings, they usually decided that something similar was present. For example, the alternative provided for the 'approved' finding 'pink' was 'red'. A 'round' border might be described as being 'oval'. When the DERMIS frequency database was reviewed in the light of these findings, it was discovered that the dermatologist had also described individual diseases using a variety of similar terms. For example, within the database 34.5% of seborrhoeic warts are described as being 'round' and 39.3% are described as being 'oval'. In this case making such a distinction will have little effect upon the posterior probability value assigned by DERMIS to the end-group seborrhoeic wart.

However, in each test case there were different sets of key features that primarily determined the ordering of the DERMIS system's differential output [Table 30].

When the diagnostic accuracy of the computer was compared with the unaided accuracy of the general practitioners, using data collected by the general practitioners, it was found that for every test case, the computer produced the correct response more frequently than the general practitioners [Table 31].

T 5 of the 8 test cases, provision of computer advice p oduced highly significant increases in the proportions of d ctors making the correct diagnosis [Table 31]. In the t ial only one general practitioner changed his diagnosis from being correct to incorrect following computer advice (a d fferent diagnosis for a benign tumour). Within the l mitations of the study, an important finding has therefore been that the provision of computer advice improved the diagnostic accuracy of practitioners viewing the test cases.

T e effects of this increase in diagnostic accuracy upon management decision making are also of great interest. For e ch case, the dermatologist had determined an 'ideal' management. However, the implications of failing to comply w'th the 'ideal ' management varied in severity between the cases. For example, case 1 was an example of a naevus. The approved' management was to reassure the patient. However, the alternative, suggested by many general practitioners, of emoving the unsightly benign lesion could be supported on social grounds. In case 5, the consequences of failing to detect and appropriately manage malignancy could well be more serious for the patient.

I cases 1 and 3, virtually all of the general practitioners correctly identified that each lesion was benign. However, t ey did not always make the correct diagnosis. Provision of mputer advice significantly increased their diagnostic ccuracy but had a smaller effect upon manageme t planning.

In cases 6, 7 and 8, patients were suffering wi h skin diseases that could adequately be treated without specialist intervention and most general practitioners in the trial offered a management plan that did not involve referral. In these cases improved accuracy lead to more frequent recommendation of appropriate medication. The was little change, in these cases, in the number of general practitioners who wished to refer the patients for expert advice. The cases are also of interest because they demonstrate that some general practitioners will refer cases that others decide to treat. This has implications for the domain definition that will be explored at a later date [5.4.a] [5.7.a].

Within the trial, the increase in general practitioner diagnostic accuracy, following computer advice, had little effect upon the total number of planned referrals [Table 31]. In general, referrals were 'saved' when benign tumours and other skin disease that could be treated in the surgery were correctly identified. More referrals occurred when malignancy was correctly identified. A conclusion from this was that the potential for 'saving' referrals [4.4] [5.4] [5.7] might only be realised when recommended management plans were included with the system.

(v) Limitations of the Semi-Field Trial

A trial involving a test set of 8 patients cannot hope to reflect the variety of skin disease presenting in the community. It was not a true field trial (167) as doctors were taken out of their normal work places and asked to pass opinions on images of patients.

However, it would be difficult to conduct a 'true' field trial of DERMIS in general practice that achieved the objectives of the semi-field trial described above. General practice does not offer the highly controlled environment of the CCU or a surgical ward. It would be difficult to establish 'gold standard' end points for disease and data c llection without disrupting the day to day workings of the practice and having a dermatologist standing by.

7 The Future Development and Testing of DERMIS

t is considered that the evidence presented supports the h pothesis that use of the DERMIS program by general practitioners is likely to improve their diagnostic accuracy and lead to the improved care of patients suffering with skin disease.

Four major areas of development and testing are planned;

a A weakness of the DERMIS database results from the collection of data from attendees at hospital outpatient clinics. A proportion of the skin disease seen in general practice, particularly in children, is transient in nature and rarely the subject of referral. For example, the database contains the records of 9 cases of chicken pox. The system was designed to assist general practitioners with difficult skin problems, but it is not known whether general practitioners will use the system on cases that they would not normally refer. A similar problem of having too few cases appl es to the rare disase groups. Priority has been given to data collection for both of these groups.

b The DERMIS system provides only diagnostic advice. It has been seen from the surveys and the sem field trial that advice [4.4] [5.4] [5.7] may also be required concerning management. Protocols of suggested management are being developed as adjunct to the system.

c) The diagnosis of skin lesions and rashes is mainly a process of pattern matching as can be seen in any dermatology clinic. The dermatologist will look at a patient and write down the diagnosis shortly afterwards. It has been considered that DERMIS could be provided with a library of images. These could be a could be called up in support of the system's differential output. In this way a general practitioner would be able to reassure himself of the final diagnosis. The images would also be useful for proving examples to match the term definitions.

d) Limited field trials will be conducted to assess patterns of usage and reactions of users. Clinical assessments will also be made on referrals to local dermatology units as as means of continuous assessment.

8. Large Domain Clinical Advice Systems: Evaluation

Within the last year, large domain expert systems such as QMR and ILIAD have started to appear in hospitals and primary care centres in the USA [1.11]. There is currently no regulation of this process and no litigation is pending. DERMIS will become part of a large domain decision support system that will be linked to an existing general practice electronic record system. Such large systems could never be effectively be subjected to double blind controlled trials in primary care. In fact the designers of some of the systems have gone as far as saying that such trials would not be appropriate (172,202).

Major problems arise through lack of domain definition and rapid development. The databases of the large systems are expanding and changing as they are kept up to date. An extensive, expensive and definitive test could be applied one month that would be invalid by the next.

When Wyatt's ACORN failed to meet its objectives of its design during field testing (25), it was not rejected but modified and retained. The same situation seems likely to arise for other systems where the users, developers or sponsors still have faith in a successful outcome. Perhaps the best that can be hoped for are the development of minimum national standards for evaluation of medical decision support systems that specify the;

- laboratory testing to be performed before implementation
- nature and frequency of assessment when implemented
- acceptable sources and frequency of update of contained knowledge
- information to be provided to the user

These could perhaps be linked to an accreditation scheme. With that in mind, I support Nykänen's view (170) that a monitored iterative development and test cycle constitutes a more realistic approach to medical advice system evaluation than Wyatt and Spiegelhalter's (167) reliance upon isolated formal trails.

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

Overview: DERMIS and Other Diagnostic Advice Systems

It is said that you don't actually learn about medicine until you start to practise it. The implication is that the study of patients can reveal more about disease than the knowledge of experts set down in books. One theme of this thesis has been the observation and negotiation of a balance between the beliefs of experts and case material as a source of knowledge for computer based decision making.

The original premise was that cases had more to offer as the implicit knowledge was less likely to have suffered interpretative bias. However as the study progressed it became obvious that expertise was requires at all stages of development of case drive systems from decisions concerning the data items to collect, definitions of terms, the diseases to study, gold standards to apply, appropriate treatments and the features that occur and do not occur in disease.

The decision to study more than one domain has at times been a burden, but has paid back rewards. In the study of acute chest pain, it was discovered that several disparate models designed to perform the same task behaved in different ways. Conclusions from this have been that choice of data set is vital and perhaps that minimalist solutions lead to loss of transferability between sites.

Within the domain of the diseases that cause acute abdominal pain, the opportunity was taken to study one of the few advice systems that has been subjected to extensive field trials. It turned out that the computer advice system had little bearing upon improvements in surgical staff performance. It was used infrequently and often too late to be of any value in decision making. Its performance did not appear to offer any diagnostic advantage to those who used it. The findings prompted the development of efficient user interfaces for DERMIS that allow decision support within the time available for consultation.

The collection of abdominal pain cases allowed investigations to be conducted into the use of the acute abdominal pain system by primary care physicians, paramedics and other personnel charged with providing medical care in remote locations. Within these groups it has been demonstrated that the acute abdominal pain system could be of value because it offers superior diagnostic performance, given adequate accurate data. The issuing of acute abdominal pain data collection sheets to general practitioners might be a simple first step that could increase their diagnostic and management performance. However, paramedics might benefit more from using the computer system, as both their sensitivity and specificity for identification of common causes of acute abdominal pain were found to be less than that of the computer.

The evolution of the DERMIS diagnostic advice system has been charted. Development decisions have been taken in the light of experience gained from both review of the literature and direct study of systems during the experimental work.

Dynamic and fixed combined frequency estimate substitution improved the diagnostic accuracy of various prototype Bayesian systems operating in the three domains. The only direct application of expert beliefs found to improve DERMIS system accuracy has been the setting of frequency estimates to zero based upon the dermatologist's identification of symptoms and signs that do not occur in diseases.

DERMIS has been designed for use by general practitioners. Components of the system include a database derived from 5203 prospectively collected clinical records, a user interface, and an enhanced Bayesian inference model incorporating combined frequency estimates, expert beliefs and rationalized end-point groups. On laboratory testing, DERMIS was able to correctly identify the diagnosis in test cases on 83% of occasions. The correct diagnosis appeared in the top three, of a possible 42 disease differential list on 97% of occasions.

In a semi-field trial of DERMIS involving 49 general practitioners, doctors did not always collect the same information as a dermatologist but were able to significantly increase their chance of making a correct diagnosis through use of the computer system. It has been concluded that although implementation of DERMIS might well increase general practitioner diagnostic accuracy and lead to improvements in the management of skin disease in primary care, rates of referral for specialist opinion might not be affected unless standard management plans are adopted.

DERMIS is set to become part of a large domain primary care advice system. Large domain systems have started to find their way into clinical use and often reside on existing hardware. Evaluation of such dynamic and extensive systems will prove difficult and should perhaps be based upon an iterative requalification procedure rather isolated definitive tests.

The Evaluation and Enhancement of Case Driven Diagnostic Advice Systems. A Study in Three Domains

List of References

1 Brahams D, Wyatt J. Decision aids and the law. Lancet 1989;ii:632-4.

2 Timpka T, Hjerppe R, Strömberg D, Möller I, et al. The need for supplements to traditional expert systems: Lessons from designing knowledge based systems for primary care. Proc. 6th International Workshop on Expert Systems, Avignon 1986.

3 de Dombal FT, Leaper DJ, Stanisland JR, et al. Computer aided diagnosis. Brit Med J 1972;2:9-12.

4 de Dombal FT, Leaper DJ, Horrocks JC. Human and computer aided diagnosis of abdominal pain. Further report with emphasis on performance of clinicians. Br Med J 1974;1:376-380.

5 Adams ID, Chan M, Clifford PC, et al. Computer aided diagnosis of acute abdominal pain: a multicentre study. Brit Med J 1986;293:800-804.

6 United Kingdom multi-centre baseline information concerning the provision of health care in acute abdominal pain provided by the National Acute Abdomen Study Association, Leeds University Medical School, Hyde Park Terrace, Leeds.

7 Mc Adam WAF, Davenport P, de Dombal. Six years experience of computer-aided diagnosis in a district general hospital. Unpublished report Airedale Hospital, Leeds 1981.

8 Clifford PC, Chan M, Hewett DJ. The acute abdomen: management with microcomputer aid. Ann Roy Coll Surg of England 1986;68:182-184.

9 Scarlett PY, Cooke WM, Clark D. Computer aided diagnosis of acute abdominal pain at Middlesbrough General Hospital. NN Roy Coll Surg of England 1986;68:179-181.

10 Hester R. Provisional medical statistics for personnel attached to nuclear powered submarines. USN Medical Report no 674, Groton 1971.

11 Health Statistics for Royal Navy Personnel provided by the Department of Statistics, Institute of Naval Medicine, Gosport, Hampshire.

12 Henderson JV, Ryack BL, Moeller G, et al. Use of a computer-aided diagnosis system aboard patrolling FMB submarines: Initial sea trials. USN Medical Report no. 938, Groton 1981.

13 Rogers W, Ryack B, Moeller G. Computer-aided diagnosis: literature review. Int J Bio-medical Computing 1979;10:267-289.

14 Osborne SF. Computer-assisted diagnosis program for acute abdominal pain: Interim report July 1982-September 1983. USN Medical report no. 1012, Groton 1983.

15 Southerland D, Fisherkeller K. ABDX- A decision support system for the management of acute abdominal pain: Users manual. USN Medical report no. 1105, Groton 1987.

16 Shaper AG, Pocock SJ, Walker M, et al. British regional heart study: cardiovascular risk factors in middle aged men in 24 towns. Br Med J 1981;282:179-186.

17 Shaper AG, Pocock SJ. Risk factors for ischaemic heart disease in British men. Br Heart J 1987;57:11-16.

18 Epstein SE, Quyyumi AA, Bonow RO. Myocardial ischaemia-silent or symptomatic. New Engl J of Med 1988;318:1038-1043.

19 Rozanski A, Bairey CN, Kranz DS, et al. Mental stress and the induction of silent myocardial ischaemia in patients with coronary artery disease. New Engl J of Med 1988;318:1005-1011.

20 Lipkin DP, Reid CJ. Myocardial infarction: the first 24 hours. Brit Med J 1988;296:947-948.

21 De Bono D. Coronary thrombolysis. Br Heart J 1987;57:301-305.

22 Van der Laarse A, Vermeer F, Hermens WT, et al. Effects of early intracoronary streptokinase on infarct size estimated from cumulative enzyme release and on enzyme release rate: a randomised trial of 533 patients with acute myocardial infarction. Am Heart J 1986;112:672-681.

23 Ikram S, Lewis S, Bucknall C, et al. Treatment of acute myocardial infarction with anisoylated plasminogen streptokinase activator complex. Br Med J 1986;293:786-89.

24 Van de Werf F, Arnold AE. Intravenous tissue plasminogen activator and size of infarct, left ventricular function, and survival in acute myocardial infarction. Brit Med J 1988;297:1374-1380.

25 Wyatt J. Lessons learnt from the field trial of ACORN, an experts system to advise on chest pain. Proc 6th World Conference on Medical Informatics Amsterdam 1989:111-115.

26 Fineberg HV, Scadden D, Goldman L. Care of patients with a low probability of acute myocardial infarction: cost effectiveness of alternatives to coronary care unit admissions. New Engl J Med 1984;310:1301-1307.

27 Zarling EJ, Sexton H, Milnor P. Failure to diagnose acute myocardial infarction. JAMA 250:1177-81.

28 Mc Queen M, Holder D, El-Maraghi N. Assessment of the accuracy of serial electrocardiography in the diagnosis of acute myocardial infarction. Am Heart J. 1893;105:258-261.

29 Poretsky L, Leibowitz IH, Friedman SA. The diagnosis of myocardial infarction by computer-derived protocol in a municipal hospital. J of Vascular Diseases 1985;3:165-170.

30 Wyatt J. The evaluation of clinical decision support systems: a discussion of the methodology used in the ACORN project. In Lecture notes in medical informatics, AIME 87, Springer-Verlag, Berlin 1987: 15-24.

31 Buxton PK. ABC of Dermatology. BMJ Books, London 1988.

32 Young DW. A survey of decision aids for clinicians. BMJ 1982;285:1332-1336.

33 Teach RL, Shortliffe EH. An analysis of physician attitudes regarding computer-based clinical consultation systems. Comput Biomed Res 1981;14:542-58.

34 Kunz JC, Shortliffe EH, Buchanan BG, Feigenbaum A. Computer-assisted decision making in medicine. J Med Phil 1984;9:135-160.

35 Bradley P. The Primary Health Care Specialist Group's view of 1992. Proceedings of the Annual Conference of the Primary Health Care Specialist Group of the British Computer Society. Sept 1992;15-20.

36 Rector AL, Nowlan WA and Kay S. Foundations for an Electronic Medical Record. Methods Inf Med 1991;30:179-86.

37 Wyatt J. Computer-based knowledge systems. Lancet 1991;338:1431-6.

38 Cannon SR, Gardner RM. Experience with a computerized interactive protocol system using HELP. Computers and Biomed Res 1980;13:399-409.

39 Wigerz O. Making decisions based on fuzzy medical data - can expert systems help? Meth Inform Med. 1986;25: 59-61.

40 Ornstein SM, Garr DR, Jenkins RG, Rust PF and Arnon A. Computer-generated physician and patient reminders. Tools to improve population adherence to selected preventive services. J Fam Pract 1991;32:82-90.

41 Classen DC, Pestotnik SL, Evans RS and Burke JP. Computerized surveillance of adverse drug events in hospital patients. JAMA 1991;226:2847-51.

42 Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959;130:9-21.

43 Lynch PJ. Skin diagnosis for the non-dermatologist. Modern Medicine Feb 1988:135-143.

44 Ashton R, Lepppard B. Differential Diagnosis in Dermatology. Radcliffe Medical Press, Oxford 1990.

45 Mc Donald CJ, Wilson GA, MCCabe GP. Physician response to computer reminders. J Amer Med Asc 1980;244:1579-1581.

46 McDonald CJ. Protocol-based computer reminders: the quality of care and the non-perfectibility of man. N Engl J Med 1976;295:1351-54.

47 Smith T. In search of consensus.(Editorial) BMJ 1991;302:800.

48 Ingram D, Bloch RF. Medical statistics: advanced techniques and computation. Mathematical Models in Medicine. John Wiley & Sons Ltd, 1984.

49 Miller RA, Pople HE, Myers JD. INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982;307:468-476.

50 Sroussi B. Computer-aided diagnosis of acute abdominal pain when taking into account interactions. Meth Inform Med 1986;25:194-198.

51 Lucas S. Neural networks and genetic algorithms: their use in medical research. Presentation to British Medical Informatics Society at U of Manchester, May 1992.

52 Evans SJW. Uses and abuses of multivariate methods in epidemiology. J Epidemiol Community Health 1988;42:311-315.

53 Joswig BC, Glover MU, Nelson DP, et al. Analysis of historical variables, risk factors and the resting electrocardiogram as an aid in the clinical diagnosis of recurrent chest pain. Comput Biol Med 1985;15:71-80.

54 Pozen MW, D'Agostino, Mitchell JB, et al. The usefulness of a predictive instrument to reduce inappropriate admissions to the coronary care unit. Ann Intern Med 1980;92:238-242.

55 Yu H, Haug PJ, Lincoln MJ, Turner C and Warner HR. Clustered knowledge representation: increasing the reliability of computerized expert systems. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. New York : IEEE Computer Society Press 1988: 126-30.

56 Faure C. Attributed strings for recognition of epileptic transients in EEG. Int J Bio-Medical Computing 1985;16:217-229.

57 Arnbjörnsson E. Scoring system for computer-aided diagnosis of acute appendicitis. Annales Chirugiae et Gynaecologiae 1985;74:159-166.

58 Fenyö G. Routine use of a scoring system for decision making in suspected acute appendicitis in adults. Acta Chir Sacnd 1987;153:545-551.

59 Hilden J. Statistical diagnosis based on conditional independence does not require it. Comput Biol Med. 1984;14:429-435.

60 Good IJ. The Estimation of Probabilities: An essay on modern Bayesian methods. MIT Press,Cambridge, Mass, 1965.

61 Bland M. An Introduction to Medical Statistics. Oxford University Press, Oxford 1991.

62 Knill-Jones RP, Stern RB, Girmes DH, et al. The use of a sequential Bayesian model in the diagnosis of jaundice by computer. Brit Med J 1973;1:530-534.

63 de Dombal FT, Pearson S, Unwin B, Mackenzie IL. Methodology in clinical decision making: Diagnostic algorithms in CAD. Proceedings of Joint international symposium on the role of non-invasive imaging modalities in clinical decision making, Leeds Sept 1985.

64 de Dombal FT. Computer-aided system for diagnosis, prognosis, and management of chest pain in a deployed situation in previously healthy adult males. US Navy Contract no. N68171/78/C/857,1979.

65 Wills KM, Teather D, Innocent PR, et al. An expert system for the diagnosis of brain tumours. Int J Man Machine Stud 1982;16:342-349.

66 Innocent PR, Teather D, Wills KM, et al. An operation system for the computer assisted diagnosis of cerebral disease. In, Van Bemmel JH, Wigertz O eds; MEDINFO 83. Amsterdam, North Holland 1983:467-470.

67 Price DJ, Salem FA. A pocket microprcessor used for the recognition of patients at risk after a head injury. In, Van Bemmel JH, Wigertz O eds; MEDINFO 83. Amsterdam, North Holland 1983:461-463.

68 Clamp SE, Myren J, Bouchier IAD, et al. Diagnosis of inflammatory bowel disease. An international multicentre scoring system. Br Med J 1982;284:91-95.

69 Ohmann C, Thon k, Stltzing H, et al. Upper gastrointestinal tract bleeding: Assessing the diagnostic contributions of the history and clinical findings.

70 Chard T. Human verses machine: a comparison of a computer "expert system" with human experts in the diagnosis of vaginal discharge. Int J Biomed Comp 1987;20:71-78.

71 Bernelot Moens HJ and Van der Korst JK. Comparison of rheumatological diagnosis by a Bayesian program and by physicians. Methods Inf Med 1991;302:935-9.

72 Carroll B. Expert systems for clinical diagnosis: are they worth the effort? Behavioural Science 1987;32:275-292.

73 Gammerman A and Thatcher AR. Bayesian diagnostic probabilities without assuming independence of symptoms. Methods Inf Med 1991;30:15-22.

74 Ohmann C, Künneke M, Zaczyk R, et al. Selection of variables using 'independence bayes' in computer-aided diagnosis of upper gastrointestinal bleeding. Stats Med 1986;5:503-515.

75 Szolovits P, Pauker SG. Categorical and probabilistic reasoning in medical diagnosis. Artificial Intelligence 1978;11:115-144.

76 Davenport PM, Morgan AG, Darnborough A, de Dombal FT. Can preliminary screening of dyspeptic patients allow more effective use of investigational techniques? Br Med J 1985;290:217-220.

77 Gear MWL, Barnes RJ. Endoscopic studies of dyspepsia in general practice. Br Med J 1980;281:1136-7.

78 Sutton GC; Simpson RJ; Holdstock G; Crichton NJ. Can preliminary screening of dyspeptic patients allow more effective use of investigational techniques? Correspondence Br Med J 1985;290:553-554.

79 Knill-Jones RP. Diagnostic systems as an aid to clinical decision making. in Logic in Medicine (Philips C ed) BMJ Pubs 1988:59-72.

80 Teather D. Statistical techniques for diagnosis. J R Statist Soc 1974;137:231-244.

81 Teather D. Diagnosis. Methods and analysis. Bull Inst Math Appl 1974;10:37-41.

82 Feinstein AR. The haze of Bayes, the aerial palaces of decision analysis, and the computerised Ouija board. Clin Pharmacol Ther 1977;21:482-496.

83 Morton BA, Teather D, du Boulay GH. Statistical Modelling and diagnostic aids. Med Decis Making 1984;3:340-348.

84 de Dombal FT. Horrocks JC, Staniland JK. Production of artificial "case histories" by using a small computer. Brit Med J 1971;11:578-581.

85 Kronmal RA, Fisher LD. The effect of assuming independence in applying Bayes' theorum to risk estimation and classification diagnosis. Comp. Biomed Res 1983;16:537-552

86 Lienert GA. Verteilungsfreie methoden in der biostatistik, Verlag Anton Hain, Meisenheim am Glan 1973.

87 Spiegelhalter DJ, Knill-Jones RP. Statistical and knowledge based approaches to decision-support systems with an application in gastroenterology. Journal of the Royal Statistical Society A 1984:147;35-77.

88 Spiegelhalter DJ. Statistical aids in clinical decision making. Statistician 1982;31:19-36.

89 Goldman L, Weinberg M, Weisberg M, et al. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. New Engl J Med 1982;307:588-596.

90 Goldman L, Cook EF, Brand DA, et al. A computer protocol to predict myocardial infarction in emergency department patients with chest pain. New Engl J Med 1988;318:797-803.

91 Goldman L. Diagnosis and management of deployed adults with chest pain. Report produced for USN contract N00014-30-C-0675 1983.

92 Zentgraf R. A note on Lancaster's definition of higher order interactions. Biometrika 1975;62:375-378.

93 Diamond GA. Computer diagnosis: revolution or revelation. Int J Cardiol 1982;2:219-220.

94 Healy MJR. Computer-aided diagnosis - an overview of some theoretical problems. Decision Making and Medical Care. North-Holland Publishing Company, Amsterdam 1976.

95 Wagner HN. Bayes' theorum: an idea whose time has come?. Am J Cardiol 1982;49:875-877.

96 Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: An Analysis of Clinical Reasoning. Harvard University Press, Cambridge, MA, 1978.

97 Norbrega GT, Morrow GW, Smoldt RK, et al. Quality assessment in hypertension: analysis of process and outcome methods. N Eng J Med 1977;296:145-148.

98 Eddy DM, Clanton CH. The art of diagnosis. New Engl J Med 1982;306:1263-1268.

99 Hershey JC, Cebul RD, Williams SV. Solid recommendations from soft numbers: The importance of considering single testing when two tests are available. Med Decis Making 1987;212-219.

100 Gorry GA. Computer-assisted clinical decision making. Methods of Information in Medicine 1973;12:45-51.

101 Dale PW. An introduction to the use of Boolean algebra to psychiatry. Psychiatric Quarterly 1955;29:48-51.

102 Ledley RS, Lusted LB. Medical diagnosis and modern decision making. Symposium in Applied Mathematics Proceedings 1961;14:117-158.

103 Feinstein AF. Clinical Judgement; Chapter 11. Williams and Wilkins, Baltimore 1967.

104 Wulff HR Rational Diagnosis and Treatment; Chapter 5. Blackwell Scientific Publications, Oxford 1967.

105 Burton JL. The logic of dermatological diagnosis. Clinical and Experimental Dermatology 1981;6:1-21.

106 Barr A, Feigenbaum EA. Handbook of Artificial Intelligence Vols 1 & 2. William Kaufman Inc, Los Altos, CA 1984.

107 Cohen PR, Feigenbaum EA. Handbook of Artificial Intelligence Vol 3. William Kaufman Inc, Los Altos, CA 1984.

108 Szolovitz P. Artificial Intelligence in Medicine. AAAS Selected symposium 51. Westview Press, 1982.

109 Clancy WJ, Shortliffe EH. Readings in Medical Artificial Intelligence: the First Decade 1983. Addison-Wesley, Reading MA.

110 Shortliffe EH. Computer-based medical consultations: MYCIN. Elsevier ,New York 1976.

111 Shortliffe EH, Scott AC, Bischoff MB, et al. ONCOCIN: An expert system for oncology protocol management. Proceedings of the American Association of Artificial Intelligence 1981:876-880.

112 Aitkins JS, Kunz JC, Shortliffe EH, Fallat RJ. PUFF: An expert system for interpretation of pulmonary function data. Computers and Biomed Res 1983; 16:199-208.

113 Hayes-Roth F, Waterman DA, Lenat DB. An overview of expert systems. In Hayes-Roth F, Waterman DA, Lenat DB,(eds). Building expert systems. Vol 1. Reading MA. Addison-Wesley Publishing Co 1983:3-29.

114 Aitkins JS. Prototypical knowledge for expert systems. Artificial intelligence 1983;20:163-210.

115 Davis R. Expert systems: where are we? and where do we go from here? The AI Magazine 1982;Spring;3-22 .

116 Wu TD. A problem decomposition method for efficient diagnosis and interpretation of multiple disorders. Computer Methods Programs Biomed 1991;35:239-50.

117 Kahn MG, Fagan LM and Tu S. Extensions to the time-orientated database model to support temporal reasoning in medcal expert systems. Methods Inf Med 1991;30:4-14.

118 Reggia JA, Perricone BT, Nau DS, et al. Answer justification in diagnostic expert systems- part II: Supporting plausible justifications. IEEE Transactions on Biomedical Engineering 1985;4:268-272.

119 Reggia JA, Peng Y. Modelling diagnostic reasoning: a summary of parsimonious covering theory. Computer Methods and Programs in Biomedicine 1987;25:125-134.

120 Gorry GA. Strategies for computer-aided diagnosis. Math Biosci. 2 1968:293-318.

121 Simon HA. The structure of ill-structured problems. Artif Intell 1973;4:181-201.

122 Warner HR. Computer-Assisted Medical Decision Making. Academic Press, New York, 1980.

123 Pryor T, Garder R, Clayton P, Warner H. The HELP system. J Med Systems 1983;7:87-102.

124 Fox J, Clark DA, Glowinski AJ, O'Neil MJ. Using predicate logic to integrate qualitative reasoning and classical decision theory. IEEE transactions on systems, man, and cybernetics 1990;20 (2): 347-357.

125 Zadeh LA. Making computers think like people. IEEE Spectrum 1984;8:26-32.

126 Hajek P. Combining functions for certainty degrees in consulting systems. Int J. Man-Machine Studies 1985;22:59-65.

127 Leaper DJ, Horrocks JC, Staniland JR, de Dombal FT. Computer assisted diagnosis of abdominal pain using "estimates" provided by clinicians. Br Med J 1972;4:350-354.

128 Haberman HF et al. DIAG: A Computer-assisted Dermatologic Diagnostic System; Clinical Experience and Insight. J Amer Acad Dermatol. 1985;12:132-143.

129 Giuse DA, Giuse NB, Bankowitz RA and Millar RA. Heuristic determination of qualitative data for knowledge acquisition in medicine.
Comput Biomed Res 1991;24:261-72.

130 Heckerling PS, Elstein AS, Terzian CG and Kushner MS. The effect of incomplete knowledge on the diagnosis of a computer consultant system. Med Inf (Lond) 1991;16:363-370.

131 Good IJ. A causal calculus (1). Brit J Philos of Science 1961;11:305-18.

132 Good IJ.A causal calculus (11). Brit J Philos of Science 1961;12:43-51.

133 Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine: where do we stand?. N Engl J Med 1987;316:685-688.

134 Cooper GF. Current research directions in the development of expert systems based on belief networks. App Stoch Models and Data Anal 1989;5:39-52.

135 Pearl J. Evidential reasoning using stochastic simulation of causal models. Artif Intell 1987;32:245-57.

136 Causal Networks: Lauritzen and Spiegelhalter
Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. J R Statist Soc B 1988;50:157-224.

137 Duran JH. Statistics and Probability. Cambridge University Press 1970.

138 Beinlich IA, Suermondt HJ, Chavez RM, Cooper GF. The ALARM monitoring system: A study of two probabilistic inference techniques for belief networks. in Proc AIME 1989, Springer-Verlag 1989: 247-256.

138 Pearl J. Fusion, propagation, and structuring in belief networks. Artif Intell 1986;29:241-88.

139 Middleton B, Shwe MA, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP and Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. II. Evaluation of diagnostic performance.
Methods Inf Med 1991;30:256-67.

140 Cooper GF. The computational complexity of probablistic inference using Baysian belief networks. Artif Intell 1990;4 393-405.

141 Chavez RM, Cooper GF. A randomized approximation algorithm for probabilistic inference on Bayesian belief networks. Networks 1990;20:661-85.

142 Herskovits EH and Cooper GF. Algorithms for Bayesian belief-network precomputation. Methods Inf Med 1991;30:81-9.

143 Hart A, Wyatt J. Connectionist models in medicine; An investigation of their potential. Proceedings AIME 1989. Springer Werlag: 115-24.

144 Whittaker J. Graphical models in applied multivariate analysis. John Wiley and Sons, Chichester 1990.

145 P Haug, PD Clayton, P Shelton, T Rich, I Tocino, PR Frederick, RO Crapo, WJ Morrison and H Warner. Revision of diagnostic logic using a clinical database. Med. Decis. Mak. 9(2);1989:84-90.

146 Fox J, Glowinski A, Gordon C, Hajnal S, O'Neil MJ. Logic engineering for knowledge engineering: design and implementation of the Oxford System of Medicine. Artificial Intelligence in Medicine 1990;2:323-339.

147 Goldberg DE. Genetic algorithms in search optimisation and machine learning. Addison Wesley, 1989.

148 Holland JH. Adaptations in natural and artificial systems. MIT Press 1992.

149 Baxt WG. Use of and artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991;115:843-8.

150 McClelland JL, Rumelhart DE. Training hidden units. In: McClelland JL, Rumelhart DE; eds. Explorations in Parallel Distributed Processing. Cambridge, Massachussetts: MIT Press; 1988:121-60.

151 Furlong JW, Dupuy ME,and Heinsimer JA. Neural network analysis of serial cardiac enzyme data. A clinical application of artificial machine intelligence. Am J Clin Pathol 1991;96:134-41.

152 Eberhart RC, Dobbins RW, Hutton LV. Neural network paradigm comparisons for appendicitis diagnosis. Proceedings of the Fourth Annual IEEE Symposium on Computer-Based Medical Systems 1991;298-304.

153 Yoon YO, Brobst RW, Bergstresser PR, Peterson LL. A desktop neural network for dermatology diagnosis. Journal of Neural Network Computation 1989;summer:43-52.

154 Bounds DG, LLoyd PJ, Mathew BG. A comparison of neural network and other pattern recognition approaches to the diagnosis of low back disorders. Neural Networks. 1990;3:583-91.

155 Hart A, Wyatt J. Evaluating "black boxes" as medical decision-aids: issues arising from a case study of neural networks. Med Informatics 1990;15:229-36.

156 Weinstein MC,Fineberg HV. Clinical decision analysis. (WB Saunders, Philadelphia 1980.

157 Inglefinger FJ. Decision in medicine. New Eng J Med 1975;293:254-255.

158 Gottleib JE, Pauker SG. Whether or not to administer amphotericin to an immuno-suppressed patient with hematologic malignancy and undiagnosed fever. Med Decis Making 1981;1:75-93.

159 Greenes RA. Computer-aided diagnostic strategy selection. Radiol Clin N Amer 1986;24:105-120.

160 Komaroff AL. The variability and inaccuracy of medical data. Proc. IEEE 1979;67:1196-1207.

161 Lehmann HP and Shortliffe EH. THOMAS: Building Bayesian statistical expert systems to aid in clinical decision making. Comput Methods Programs Biomed 1991;35:251-60.

162 Reggia JA, Perricone BT. Answer justification in medical decision support systems based on bayesian classification. Comp Biol Med 1985;15:161-167.

163 Hilden J, Habbema JDF, Bjerregaard. The measurement of performance in probabilistic diagnosis III: Methods based on continuous functions of the diagnostic probabilities. Meth Inform Med. 1978;17:238.

164 Miller RA, Masarie FE. The demise of the Greek oracle model for medical diagnosis systems. Methods Inf Med 1990;29:1-2.

165 Charniak E. The Bayesian basis of common sense medical diagnosis. Proceedings of the National Conference on Artificial Intelligence, AAAI. William-Kaufman Inc. 1983:70-73.

166 Van der Lei J, Musen MA, Van der Does E, Man in't Veld AJ and Van Bemmel JH. Comparison of computer-aided and human review of general practitioners' management of hypertension. Lancet 1991;338:1504-8.

167 Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Med Inform 1990;3:205-217.

168 Rosati RA, Wallace AG, Stead EA. The way of the future. Archives of Internal Medicine 1973;131:285-287.

169 Lundsgaarde HP. Evaluating medical expert systems. Social Science and Medicine 1987;24:805-19.

170 Nykänen P,Chowdhury S and Wigertz O. Evaluation of decision support systems in medicine. Comput Methods Programs Biomed 1991;34:229-38.

171 Schreiner A, Chard T. Expert systems for the prediction of ovulation: comparison of an expert shell (Expertec Xi Plus) with a program written in a traditional language (BASIC). Methods of Information in Medicine 1990;29 (2):140-145.

172 Feldman MJ and Barnett GO. An approach to evaluating the accuracy of DXplain. Comput Methods Programs Biomed 1991;35:261-6.

173 Gill PW, Leaper DJ, Guillou PJ, Staniland JR, Horrocks JC, and de Dombal FT. Observer variation in clinical diagnosis- A computer aided assessment of its magnitude and importance in 552 patients with abdominal pain. Methods Inform. Med. 12;1973:108-113.

174 de Dombal FT. Analysis of symptoms in the acute abdomen. Clin Gastroenterol 1985;14:531-43.

175 Card WI, Lucas RW, Spiegelhalter DJ. The logical description of a disease class as a Boolean function with special reference to the irritable bowel syndrome. Clinical Science 1984;66:307-315.

176 Titterington DM, Murray GD, Murray LS, et al. Comparison of discrimination techniques applied to a complex data set of head injured patients. J R Statist Soc 1981;144:145-175.

177 Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodological standards. N Engl J Med 1985;313:793-99.

178 Zentgraf R, Victor N. Some problems arising in the statistical treatment of diagnosis. Meth Inoform Med 1978;17:10-15.

179 Franklin RGC, Spiegelhalter DJ, Macartney FJ, and Bull K. Evaluation of a diagnostic algorithm for heart disease in neonates. Br Med J 1991;302:935-9.

180 Brooks GJ, Ashton RE, Pethybridge RJ. DERMIS: a computer system for assisting primary care physicians with dermatology diagnosis. British Journal of Dermatology 1992;127:614-619.

181 Alvey P, Greaves MF. Observations on the development of a high performance system for leukaemia diagnosis; in Proc. Expert Systems '86 (Brighton), ed. Bramer M.,Cambridge University Press 1986.

182 Puppe B, Puppe F. MED1: An intelligent computer program for thoracic pain diagnosis. In Klinische Wochen-schriff, Springer-Verlag, Berlin 1985:511-517.

183 Patrick EA, Moskowitz M, Mansukhani VT and Gruenstein EI. Expert learning system network for diagnosis of breast calcifications. invest Radiol 1991;26:534-9.

184 Pople HE. CADUCEUS: An experimental expert system for medical diagnosis. In Winston PH, Prendergast KA (eds). The AI business:The commercial use of artificial intelligence. Cambridge MA: The MIT Press 1984:67-80.

185 Yu VL. Antimicrobial selection by a computer. A blinded evaluation by infectious disease experts. JAMA 1979;242:1279-1282.

186 Yu VL, Fagan LM, Wraith SM, et al., Antimicrobial selection by a computer: blinded evaluation by infectious disease experts. J Am. Med. Assoc. 242;1979:1279-1282.

187 Lundsgaarde HP. Evaluating medical expert systems. Soc. Sci. Med. 24(10);1987:805-819.

188 de Dombal FT, Dalos V and Mc Adam WAF. Can computer aided teaching packages improve clincal care in patients with acute abdominal pain? Br Med J 1991;302:1495-7.

189 Startsman TS, Robinson RE. The attitudes of medical and paramedical personnel towards computers. Computers and Biomed Res 1972;5:218-227.

190 Shortliffe EH. Testing Reality. The introduction of decision support technologies for physicians. Methods Inform. Med. 28;1989:1-5.

191 Bankowitz RA, Miller Mc Neil MA, Challinor SM, Miller RA. Effect of a computer-assisted general medicine consultation service on housestaff diagnostic strategy. Meth Inf Med 1989;28:352-6.

192 Kassirer JP. Our stubborn quest for diagnostic certainty. A cause of excessive testing. N Engl J Medicine 1989;320:1489-91.

193 J Gaschnig, P Klahr, H Pople, E Shortliffe and A Terry. Evaluation of expert systems: issues and case studies, in: Building Expert Systems, eds. F Hayes-Roth, DA Waterman and DB Lenat, pp 241-280 (Addison-Wesley, New York, 1983).

194 Ginzler M, Pritchrd PMM. "Can medical knowledge-based system cross frontiers?", LEMMA project deliverable, Advanced Informatics in Medicine (AIM) programme for the European Commission, 1990.

195 Spiegelhalter DJ. Evaluation of clinical decision-aids, with an application to a system for dyspepsia. Stats in Medicine 1983;2:207-216.

196 Nelson SJ, Blois MS, Tuttle MS, Erlbaum M, Harrison P, Kim H, Winkelmann B, and Yamashita D. Evaluating RECONSIDER. A computer program for diagnostic prompting. Journal of Medical Systems 1985;9:379-388.

197 Waxman HS, and Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new micro-computer based system. Medicine 1990;69:125-136.

198 Barnet GO, Cimino JJ, Hupp JA, and Hoffer EP. DXplain - an evolving diagnostic decision support system. Journal of the American Medical Association 1990;258:67-74.

199 Miller RA, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Computing 1986;3:34-38.

200 Warner HR Jr. Iliad: moving medical decision making into new frontiers. Methods of Information in Medicine 1989;28:370-2.

201 Fox J, Ginzler M, Glowinski A, et al. Technicalities and practicalities of logic engineering in medicine: the LEMMA project.LEMMA project deliverable, Advanced Informatics in Medicine (AIM) programme for the European Commission, 1990.

202 Fox J, Glowinski A, Gordon C, Hajnal S, O'Neil M. Logic engineering for knowledge engineering: design and implementation of the Oxford System of Medicine. Artif Intell Med. 1990;2:323-39.

203 Weiner, Laufer D, Ribak A. Computer-aided diagnosis of odontogenic lesions. Int J Oral Maxillofac Surg 1986;15:592-596.

204 Hagen MD. A pocket calculator program for using Pozen's formula. Comput Biol Med 1986;16:155-157.

205 Hall PN. Computer aided diagnosis of acute abdominal pain. Brit Med J 1986;293:1025

206 de Dombal FT. How to interview and examine patients suffering from acute abdominal pain. OMGE pamphlet 1986.

207 Smith SB, Reyna TM, Hollis HW. Periappendicitis: possible surgical pitfall. Military Medicine 1986;151:612-613

208 Dixon JAS, Edwards J, Jones AP. Computer-assisted diagnosis of acute abdominal pain in childhood. The Lancet 1985:1389-1390

209 Stoeker WV. Computer-Aided Diagnosis of Dermatologic Disorders. Dermatologic Clinics. 1986;4:607-625

210 Chard T. Self-learning for a Bayesian knowledge base: How long does it take for the machine to educate itself? Meth Inform Med 1987;26:185-188

211 Dolan JG, Bordley DR, Mushlin AI. An evaluation of clinicians' subjective prior probability estimates. Med Decision Making 1986;6:16-223

212 Fisherkeller, Moeller G, Ryack BL, Goudy J. An evaluation of the ability of Navy hospital corpsmen to collect chest pain data from patients. USN Medical report no. 1016, Groton 1984.

213 Bailey NTJ. Probability methods of diagnosis based on small samples. In Mathematics and Computer Science in Biology and Medicine. London 1984;HMSO:103-110

214 Wiegert HT, Weigert OA. Acute ischaemic heart disease: an empirical trial in the use of the Pozen and D'Agostino formula. J Kentucky Med Assoc 1983;11:834-836

215 Willems JL,Abreu-Lima C,Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. N Engl J Med 1991;325:1767-73

216 Pozen MW, D'Agostino RB, Selker HP, et al. A predictive instrument to improve coronary care unit admission practices in acute ischaemic heart disease. New Engl J Med 1984;310:1274-1278

217 Brooks GJ. Royal Navy acute abdominal pain diagnostic assistant. INM Technical Memo 6/87, Gosport, Hampshire.

218 Good IJ. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Press, Cambridge, Massachusetts 1965.