Test-selection Strategies for Probabilistic Networks

Testselectie-strategieën voor probabilistische netwerken

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de Rector Magnificus, Prof. dr. W.H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op maandag 17 oktober 2005 des middags te 14:30 uur

door

Daniëlle Sent

geboren op 21 november 1977, te Rheden

promotor: Prof. dr. ir. L.C. van der Gaag

Faculteit Bètawetenschappen, Departement Informatica, Universiteit Utrecht

beoordelingscommissie:

Dr. S.K. Andersen (Aalborg University)
Prof. dr. Y. van der Graaf (Universiteit Utrecht)
Prof. dr. J.-J.Ch. Meyer (Universiteit Utrecht)
Dr. S. Quaglini (University of Pavia)
Prof. dr. A.P.J.M. Siebes (Universiteit Utrecht)

The research reported in this thesis was supported by:

NWO
Nederlandse Organisatie voor Wetenschappelijk Onderzoek

SIKS Dissertation Series No. 2005 – 18

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Graduate School for Information and Knowledge Systems.

ISBN-10: 90393340234
ISBN-13: 97890393340233

Contents

CHAPTER 1

Introduction

A decision-support system is a computer-based system that provides support to human decision makers in their problem-solving processes. The concept of decision support has evolved from two main research areas: mathematical decision making and computer science. An effective decision-support system contains a reliable up-to-date knowledge base, accurate data, a user-friendly interface, and a sound inference mechanism. The knowledge base captures the knowledge of the application domain that is required to provide support at an expert level. The inference mechanism applies the knowledge for making decisions. Decision-support systems typically are designed to deal with problems that require considerable human expert knowledge for their solution and are being developed for various different domains of application such as weather forecasting, predicting the oil market, deciding whether to launch a new product, diagnosing a technical problem in air conditioning systems, and responding to environmental risks [7, 27, 67].

In this thesis we will focus on decision support in the field of medicine. Since over the last decades researchers have come to understand more and more of diseases and their management, it nowadays is hard, even for medical specialists, to keep up-to-date with medical literature and with new insights of diseases, new drugs and new procedures. On the other hand, the costs of more sophisticated treatments, tests and procedures are increasing. Hospitals have to provide care with a limited budget and yet have to keep care at the highest possible level. Furthermore, physicians have to be more and more aware that any mistake they make, not only will affect the patient under their care, but can also result in professional sanctions and financial compensation. To cope with the increasing complexity of medical practice and to provide for a constant level of care, medical procedures are becoming more and more standardised. For example, various different medical protocols and guidelines have been developed in which the physicians' daily problem-solving strategies have been caught [16, 46, 55]. Decision-support systems can also assist physicians in their complex problem-solving tasks, by providing support that is tailored to

individual patients.

To be able to provide actual support, a decision-support system should first of all gain the confidence of the physicians and prove to be of added value to their daily practice, for example by resulting in lower costs, shorter waiting lists, less discomfort for patients and better care. The conclusions that are reached by the system should be correct according to the state of the art in medical practice; furthermore, these conclusions should really *support* its users, rather than, for instance, only provide information they could easily have reached themselves. At every step in the decision-making process, the system should further be able to provide information about why it has reached particular conclusions and why it has rejected other conclusions. The system should thus be able to explain why a specific disease is a likely explanation of a patient's findings; it should also be able, on the other hand, to explain why other diseases have been eliminated as possible conclusions. The decision-support system should further allow for studying different *what-if* scenario's. It should be able to provide information, for instance, about the most likely disease if a test result would have been different. To allow for easy usage, it should further be integrated with the other computer systems employed providing links to glossaries and medical literature and a connection with electronic patient records.

To support the entire process of a patient's management, a decision-support system should not only provide information about the most probable diseases or the best suitable therapy, it should also provide its user with information about which diagnostic tests should be performed. These diagnostic tests then are performed to reduce the uncertainty about a patient's true condition. In this thesis we will focus on test selection for decision-support systems. More specifically we will focus on test selection for systems that employ a probabilistic network for its reasoning tasks. Decision-support systems may in general be based upon such techniques as classification trees, decision trees, neural networks, rule-based systems, and probabilistic networks. Decision-support systems in the field of medicine, where reasoning with uncertainty is quite prominent, are increasingly based upon a probabilistic network. A *probabilistic network* is a concise representation of a joint probability distribution on a set of statistical variables. It consists of a directed acyclic graph and an associated set of numerical parameters. Each node in the graph represents a variable that can take a value from among a set of mutually exclusive discrete values. The arcs in the graph represent direct influential relationships between the variables. Associated with the graph of the network are conditional probabilities that describe the strengths of the relationships between the variables [10, 28, 44]. Various different probabilistic networks have been developed in the field of medicine. Examples include the MUNIN network for the interpretation of electromyographic findings [2, 3], the QMR network for diagnosis in internal medicine [40, 57], the DIAVAL network for the diagnosis of heart diseases [12], the Pathfinder network that assists pathologists with the diagnosis of lymph node diseases [26], and many others [34, 37, 43]; at Utrecht University, a network-based decision-support system has been constructed for the domain of oesophageal cancer [61, 62].

The aim of this thesis is to develop a test-selection facility for decision-support systems that include a probabilistic network. The facility should induce a test-selection strategy that is based upon the mathematical principles of decision theory, yet closely fits the test-selection strategy employed by physicians in their daily practice. To study the practicability of the concepts and algorithms involved, we evaluate the use of our new facility in the context of the oesophageal

cancer network.

The oesophageal cancer network that we will use to illustrate our concepts, provides for the staging of cancer of the oesophagus. It models the presentation properties of an oesophageal tumour, such as its length and macroscopic shape, and describes the pathophysiological processes that influence its growth and metastasis. The network further represents the various diagnostic tests that are commonly used to gain insight in the properties of the tumour and the extent of its advance. The framework of probabilistic networks as well as the oesophageal cancer network more specifically, are described in Chapter 2.

An automated test-selection facility thus consists of three basic elements: a measure for establishing the usefulness of performing a particular test, a test-selection loop, and a criterion for deciding when to stop gathering further information. Many researchers describe a myopic test-selection facility. A myopic facility will select a single test and prompt the user for its results. After processing the result, it selects the next test. In non-myopic test selection, other than in myopic test selection, not the most informative test, but rather the most informative combination of tests is selected. A non-myopic test-selection facility offers more flexibility in test selection than a myopic one, yet brings a high computational burden. In Chapter 3 we discuss the state of the art of the basic elements of test selection.

Diagnostic tests generally do not unambiguously reveal the true condition of a patient. To avoid misdiagnosis, therefore, the reliability characteristics of the various tests employed should be taken into consideration upon constructing a diagnostic hypothesis. Diagnostic tests are often characterised by their sensitivity, that is, the probability that a test result is positive when the underlying pathological state tested for is present, and their specificity, that is, the probability that a test result is negative when the underlying pathological state tested for indeed is absent. These standard concepts are described for binary disease variables, with the values present and absent, and binary test variables, with the values positive and negative. A probabilistic network, however, may include non-binary disease variables as well as non-binary test variables for which more detailed characteristics are required. The network may further require these characteristics to be stratified over several different subgroups of patients. To provide for including test characteristics into a probabilistic network, therefore, we extended the standard concepts of sensitivity and specificity. The modelling of generalised reliability characteristics in probabilistic networks is discussed in Chapter 4.

Building upon a probabilistic network's mathematical foundation, a myopic test-selection strategy could easily be designed. Such a strategy would select, on a one-by-one basis, the test variable that is the most informative given the already available evidence [1, 13]. To compute the most informative test, often an information measure is used. The two most commonly used measures for capturing diagnostic uncertainty in decision-support systems, are the Shannon entropy and the Gini index [19]; in other contexts, also the misclassification error is used for measuring uncertainty [24]. The three measures are defined for a probability distribution over a designated variable and express the expected amount of information that is required to establish the value of this variable with certainty. The Shannon entropy and the Gini index are generally considered to behave alike for test-selection purposes, in particular for variables with a small number of values [5]. In fact, common knowledge has it that the two measures are interchangeable in practice. In addition, the Shannon entropy and the Gini index are commonly acknowledged to be

more sensitive to changes in the probability distribution over the main variable of interest than the misclassification error [6]. In Chapter 5 we compare the Shannon entropy, the Gini index and the misclassification error both from a fundamental and an experimental perspective.

Upon working with the oesophageal cancer network, we felt that a myopic test-selection strategy would be an oversimplification of the experts' problem-solving practice. To acquire knowledge about the actual test-selection strategy employed by our experts and about the arguments underlying their strategy more specifically, we designed an elicitation method that consists of three main interviews: an unstructured interview, followed up by a structured one and an interview for refinement purposes. The aim of the first interview was to elicit general knowledge about the selection of diagnostic tests for patients suffering from oesophageal cancer. The goal of the second interview was to fill in the details of the general test-selection strategy that had been acquired from the two experts during the first interview. More specifically, we wanted to elicit the exact arguments used by the experts in their decisions to order new tests or to refrain from further testing. During the third and last interview, we refined the elicited test-selection strategy. Since we would be using the arguments underlying our experts' problem-solving practice to build a facility for automated test selection, we had to ascertain that we had elicited their test-selection strategy at a sufficient level of detail. To this end, we studied the medical records of real patients and compared these against the flow chart that we had constructed after the third interview. The results from the interviews and from the analysis of the patient records, are described in Chapter 6.

From the three interviews that we held with our experts in the domain of oesophageal cancer, we learned that in their daily test-selection routines they focused on different subgoals that are addressed sequentially. We felt that a more involved test-selection facility should be able to take such subgoals into account. We also felt that the myopic nature of the algorithms currently in use presents an oversimplification of test selection in many fields in medicine. In the domain of oesophageal cancer, for example, multiple test variables serve to model the results of a single physical test. Moreover, our experts have been found to order tests in packages to reduce the length in time of the diagnostic phase of a patient's management. A non-myopic test-selection algorithm would thus be indicated. In Chapter 7, we present new algorithms for automated test selection. The three algorithms are all enhanced with a sequence of subgoals to be addressed. The algorithms differ mainly in their degree of non-myopia and range from fully myopic, via semi-myopic to non-myopic. To prevent overtesting, the test-selection algorithms are extended with a stopping criterion. With such a criterion, the algorithm decides in every step of the test-selection process if performing more diagnostic tests is necessary or the physician can safely stop testing. We designed four stopping criteria and will discuss these from a fundamental as well as from a experimental perspective.

This thesis ends with our conclusions and suggestions for future research in Chapter 8.

CHAPTER 2

Probabilistic networks

During the mid 1980s, probabilistic networks were introduced as models for reasoning with uncertainty in decision-support systems [44]. In this chapter we will briefly review the framework of probabilistic networks and describe the oesophageal cancer network that has been developed at Utrecht University over the past decade. The oesophageal cancer network will be used as an example throughout this thesis.

2.1 Probabilistic networks

Probabilistic networks, also known as (Bayesian) belief networks, Bayes nets, or causal networks [10, 28, 44], are graphical models of uncertainty. They are most suitable for representing knowledge in domains in which uncertainty is predominant and are being used in decision-support systems in a wide range of applications such as in the medical domain. In this section we briefly review the framework of probabilistic networks.

A *probabilistic network* is a concise representation of a joint probability distribution $\Pr$ on a set of statistical variables. It consists of a directed acyclic graph and an associated set of numerical parameters. Each node in the graph represents a variable that can take a value from among a set of mutually exclusive discrete values. Variables will be indicated by capital letters, for example A; the possible values of A are denoted $a_1, \ldots, a_n$, $n \geq 1$. The arcs in the graph represent direct influential relationships between the variables. Informally, an arc $A \rightarrow B$ can often be interpreted as expressing that the variable A has a causal influence on the variable B. More formally, the set of arcs captures probabilistic independence: two variables are said to be independent given the available observations if every chain between the two variables contains an observed variable with at least one emanating arc, or a variable with two incoming arcs such that neither the variable itself nor any of its descendants in the graph have been observed. Associated

with the graph of the network are conditional probability distributions that describe the strengths of the relationships between the variables. More specifically, for each variable A are specified the conditional probabilities $p(A \mid \pi(A))$ that describe the joint influence of the various values of the variable's parents $\pi(A)$ on the probabilities of the values of A itself. Figure 2.1 shows, as an example, a small fragment of a probabilistic network in the field of oesophageal cancer. The graphical structure from Figure 2.1(a) expresses that the location of a primary tumour in the oesophagus influences the histological type of its cells. The table from Figure 2.1(b) further shows the nine conditional probabilities that are specified for the variable *Type*; it states for example that $p(\textit{Type} = \text{adeno} \mid \textit{Location} = \text{distal}) = 0.85$, which serves to indicate that a tumour that is located in the lower, or distal, region of the oesophagus, is most likely to be composed of adenous cells.

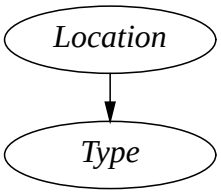
The graphical structure of a probabilistic network and its associated probabilities with each other uniquely define a joint probability distribution $\Pr$. A probabilistic network thereby provides for computing any probability of interest over its variables. In the sequel, we will explicitly distinguish between computed probabilities, for example $\Pr(A)$, and the parameter probabilities $p(A \mid \pi(A))$ that are specified in the network. To illustrate the computation of probabilities of interest, we consider again the network fragment from Figure 2.1. The prior probability of a squamous-cell tumour, for example, can be computed from the probabilities specified for the fragment by means of the conditioning rule of probability:

$$\begin{aligned}
&\Pr(\textit{Type} = \text{squamous}) = \\
&\quad = p(\textit{Type} = \text{squamous} \mid \textit{Location} = \text{proximal}) \cdot p(\textit{Location} = \text{proximal}) + \\
&\qquad p(\textit{Type} = \text{squamous} \mid \textit{Location} = \text{mid}) \cdot p(\textit{Location} = \text{mid}) + \\
&\qquad p(\textit{Type} = \text{squamous} \mid \textit{Location} = \text{distal}) \cdot p(\textit{Location} = \text{distal}) = \\
&\quad = 0.84 \cdot 0.10 + 0.64 \cdot 0.40 + 0.14 \cdot 0.50 = 0.41
\end{aligned}$$

The probability of the primary tumour being located in the mid region of the oesophagus given that its cells are squamous cells, can be computed using Bayes' rule:

$$\begin{aligned}
&\Pr(\textit{Location} = \text{mid} \mid \textit{Type} = \text{squamous}) = \\
&\qquad = \frac{p(\textit{Type} = \text{squamous} \mid \textit{Location} = \text{mid}) \cdot p(\textit{Location} = \text{mid})}{\Pr(\textit{Type} = \text{squamous})} = \\
&\qquad = \frac{0.64 \cdot 0.40}{0.41} = 0.62
\end{aligned}$$

For computing probabilities of interest, various different inference algorithms for *probabilistic inference* are available [36, 44]. These algorithms employ the graphical structure of a network more or less directly as a computing architecture. Evidence that has been entered for the variables, is propagated throughout the network, thereby updating the probability distribution over all the variables. Exact probabilistic inference is known to be NP-hard in general [9]. The available algorithms thus have a computational complexity that is exponential in the size of a probabilistic network in general. For relatively sparse networks from which probability distributions per variable have to established, however, probabilistic inference tends to be feasible for practical purposes.



(a)

		Location		
		proximal	mid	distal
Type	adeno	0.15	0.35	0.85
	squamous	0.84	0.64	0.14
	undiff	0.01	0.01	0.01

(b)

Figure 2.1: A small fragment of a probabilistic network in oncology.

2.2 The oesophageal cancer network

The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, is a highly specialised center for the treatment of patients suffering from cancer in the Netherlands. Some eighty patients receive treatment for oesophageal cancer at this center per year. These patients receive diagnostic tests and therapies based upon a standardised protocol. Over the past decade, a decision-support system has been developed that is aimed at a better tailored patient-specific management of oesophageal cancer. This system has a probabilistic network at its kernel, that has been constructed and refined with the help of two experts in gastrointestinal oncology from the Netherlands Cancer Institute. We will refer to this network as the *oesophageal cancer network*. In this thesis, we will use the oesophageal cancer network as an example; the fragment shown in Figure 2.1 in fact was taken from this network. The network currently includes 42 variables, for which almost 1000 parameter probabilities have been specified [62, 63].

An oesophageal tumour may develop within the oesophagus as a consequence of a lesion of the oesophageal wall, for example as a result of chronic reflux, the intake of spicy food or associated with smoking and drinking habits. Due to the presence of an oesophageal tumour, a patient will often suffer from difficulties with swallowing food and, consequently, from weightloss [42]. A patient's ability to swallow food depends on the characteristics of the primary tumour, such as its location in the oesophagus, its histological type, shape and length. These characteristics also influence the tumour's prospective growth. The histological type is established from a biopsy, during which a small amount of tissue is taken from the primary tumour. The other characteristics of the tumour are investigated during a gastroscopic examination. Upon the examination, a camera attached to a flexible tube is brought into the oesophagus; the images of the camera show the oesophageal wall and the primary tumour. Note that if the tumour has developed a considerable mass that protrudes into the oesophagus, the camera may not be able to pass by it.

An oesophageal tumour typically invades the oesophageal wall upon growth. Once it has grown through all three layers of the oesophageal wall, it may also invade neighbouring organs

such as the trachea and bronchi, the heart, the mediastinum, or the diaphragm. Which neighbouring organs are invaded depends on the location of the tumour in the oesophagus. The depth of invasion into the oesophageal wall or into neighbouring organs is investigated with a bronchoscopy, an endosonographic examination, a barium swallow, a CT-scan of the patient's thorax, and a laparoscopic examination. A bronchoscopy is an endoscopic examination with a flexible tube equipped with a camera that is brought into the bronchi of the patient; images from the camera provide information about the absence or presence of invasion of the primary tumour into the bronchi. For a barium swallow, the patient swallows a fluid containing barium. Then, an X-ray is taken, from which the physician can establish, for example, the location of the tumour and the passage of fluids. The physician can also decide from the X-ray whether or not a fistula is present. During an endosonography, a flexible tube with a sonographic device is brought into the patient's oesophagus. The sonographic device uses high-frequency sound waves to produce images of the organs and structures in the patient's body. From these images, for example, the various layers of the oesophageal wall can be examined. A laparoscopic examination is in essence a surgical procedure. During this procedure a non-flexible tube with a camera is inserted through the patient's skin to the location of interest. The images from the camera provide information about, for instance, the invasion of the primary tumour into the patient's diaphragm.

The primary tumour may give rise to secondary tumours. Secondary tumours, or metastases, can be either lymphatic or haematogenous. Lymphatic metastases develop in lymphatic nodes from tumour cells having been conveyed through the lymphatic vessels. Haematogenous metastases, on the other hand, develop as a consequence of tumour cells having been conveyed through the patient's blood vessels. Lymphatic metastases are typically located in the lymph nodes in the patient's neck or in the lymph nodes near the truncus coelicus. The lymph nodes that will be affected first, however, are the loco-regional lymph nodes located near the primary tumour. The presence or absence of lymphatic metastases is investigated with an endo-sonography of the oesophagus, a laparoscopy, a sonography of the patient's neck, and a physical examination during which the neck is examined for palpable lymphatic nodes. Haematogenous metastases are typically located in the lungs and the liver of the patient. The presence or absence of haematogenous metastases is established from a laparoscopic examination, a CT-scan of the thorax and of the abdomen, and an X-ray of the patient's lungs.

The depth of invasion of the primary tumour into the oesophageal wall and the extent of the different types of metastasis are summarised in the *stage*, or *TNM classification*, of the cancer. In this classification, the T class represents the extent of invasion into the oesophageal wall by the primary tumour. The N class summarises the presence or absence of lymphatic metastases, and the M class denotes the presence or absence of metastases distant from the primary tumour. The T class can take the values $T1, \ldots, T4$, where $T4$ indicates that the tumour has grown through all layers of the oesophageal wall. The N class is either $N0$ or $N1$, indicating either the absence or the presence of lymphatic metastases. $M0$ and $M1$ indicate either the absence or the presence of metastases distant from the primary tumour. All possible combinations of the classes T, N and M are summarised in the six possible stages, *I, IIA, IIB, III, IVA*, and *IVB*. The stage of the cancer together with the patient's physical condition, largely influence a patient's life expectancy and are indicative of the effects and complications to be expected from the various different therapeutic alternatives.

The oesophageal cancer network in essence provides for the staging of cancer of the oesophagus. It models the presentation characteristics of an oesophageal tumour and describes the pathophysiological processes that influence its growth and metastasis. The network further represents the various diagnostic tests that are commonly used to gain insight in the properties of the tumour and the extent of its advance. The main diagnostic variable of the network is the variable *Stage* that captures the extent to which the cancer has progressed. Of the 42 variables included in the network, 25 variables serve to model the results of the 12 physical tests described above. Figure 2.2 depicts the graphical structure of the network and the prior probabilities per variable.

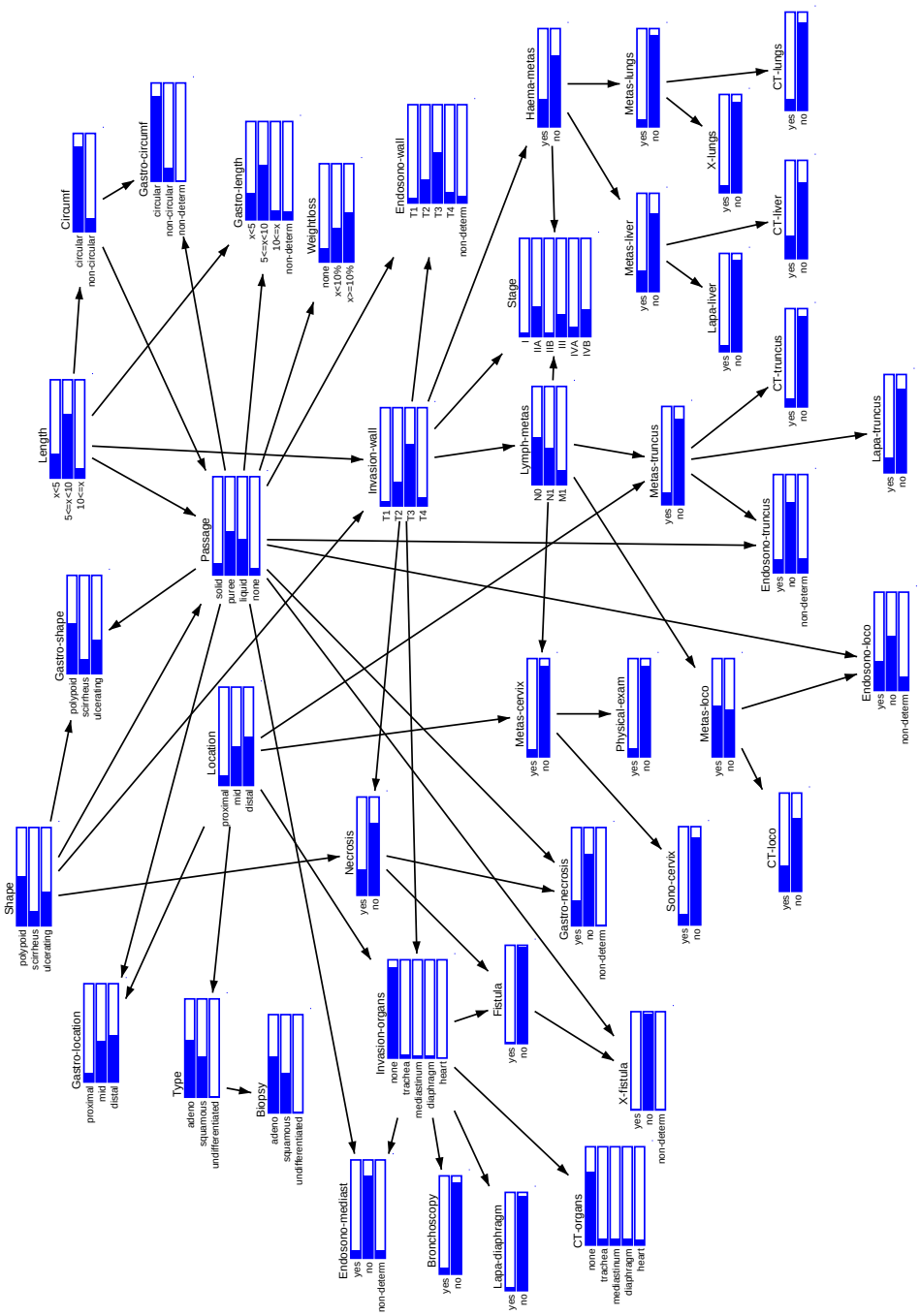


Figure 2.2: The oesophageal cancer network.

CHAPTER 3

Diagnostic tests and test selection

In this chapter we briefly review the various different concepts and techniques that are commonly used in automated test selection. In Section 3.1, we will focus on basic concepts that originate from the field of medical decision making. In Section 3.2 we will address the state of the art in automated test selection.

3.1 Concepts in test selection

An important issue in test selection is how to value the information that is gathered from tests. Test results are not always unequivocal, since the tests involved generally do not unambiguously reveal the true condition tested for. To avoid misdiagnosis, therefore, the reliability characteristics of the tests employed should be taken into consideration upon constructing a diagnostic hypothesis [17, 58]. The information value of the results of diagnostic tests is studied especially in the field of medical decision making. In this field, specialised concepts have been developed to capture the information value of a test result. Since we will often use the information characteristics of diagnostic tests in this thesis, we will briefly review these concepts.

3.1.1 Sensitivity and specificity

The standard concepts of sensitivity and specificity have been designed to capture the uncertainty in the results of a diagnostic test, and thereby constitute the test's reliability characteristics [17, 23, 58, 66]. The concepts are defined in terms of two binary variables. We use the variable D to represent the presence, indicated by the value d, or absence, indicated by $\overline{d}$, of a specific disease; D will often be referred to as the *disease variable*. The variable T describes the result of an associated diagnostic test, where the value t models a *positive* test result, suggesting presence of

the disease, and the value $\bar{t}$ models a *negative* result, suggesting absence of the disease; T will be termed the *test variable*.

The *sensitivity*, or *true-positive rate* (TPR), of a test now is defined as the probability that a positive test result is found in a patient who actually has the disease; more formally, it is defined as $\Pr(t \mid d)$. The complement $\Pr(\bar{t} \mid d)$ of the sensitivity is the probability that a negative result is found in a patient who has the disease tested for; the complement of a test's sensitivity is sometimes called the *false-negative rate* (FNR) of the test. Note that, since $\Pr(t \mid d) + \Pr(\bar{t} \mid d) = 1$, we have that the true-positive rate and the false-negative rate of a test sum to one. The *specificity*, or *true-negative rate* (TNR), of a test is the probability $\Pr(\bar{t} \mid \bar{d})$ that the test yields a negative result for a patient without the disease. The complement $\Pr(t \mid \bar{d})$ of the specificity is the probability that a patient without the disease shows a positive test result; the complement of the specificity is sometimes called the *false-positive rate* (FPR) of the test. Note that, since $\Pr(\bar{t} \mid \bar{d}) + \Pr(t \mid \bar{d}) = 1$, we have that the true-negative rate and the false-positive rate sum to one. We further note that a very sensitive test is good at detecting patients with the disease under consideration, while a very specific test is good at singling out patients who do not have the disease. The two concepts therefore pertain to different groups of patients and, hence, measure truly different reliability characteristics of a diagnostic test [58].

In the remainder of this thesis, we will often summarise the reliability characteristics of a test in a 2×2 *probability table*. As an example, Table 3.1 presents such a table for the relationship between the true presence or absence of secondary tumours in the liver of a patient, captured by the disease variable *Metas-liver*, and the result of a CT-scan of the patient's liver, captured by the test variable *CT-liver*. The table describes the relationship in terms of the four probabilities: $\Pr(\textit{CT-liver} = \text{yes} \mid \textit{Metas-liver} = \text{yes})$ which represents the sensitivity of the scan, $\Pr(\textit{CT-liver} = \text{yes} \mid \textit{Metas-liver} = \text{no})$ which represents its false-positive rate, $\Pr(\textit{CT-liver} = \text{no} \mid \textit{Metas-liver} = \text{yes})$ which captures its false-negative rate, and $\Pr(\textit{CT-liver} = \text{no} \mid \textit{Metas-liver} = \text{no})$ which represents the specificity of the CT scan for establishing the absence or presence of secondary tumours in a patient's liver.

The sensitivity and specificity of diagnostic tests are commonly established from the results of a clinical study. In such a study, a population of individuals is tested. For each individual it is known whether or not he has the disease tested for. For each individual, moreover, the result of the test under study is established. The outcomes of such a clinical study are summarised in a 2×2 *frequency table*. Such a table differs from a 2×2 probability table in that it captures numbers of individuals rather than probabilities. Table 3.2 presents such a 2×2 frequency table from a (fictitious) study of the reliability characteristics of a CT-scan of a patient's liver. The individuals under study with a positive test result who have the disease tested for are called the *true positives*. The *false negatives* are the individuals with a negative test result who do have the disease under study. The *false positives* are the individuals in whom the disease tested for is absent, although their test result is positive. The individuals without the disease who have a negative test result constitute the *false negatives*. The sensitivity (TPR) of the test can now be readily computed from the frequency table by dividing the number of true positives (TP) by the

sum of the number of true positives (TP) and the number of false negatives (FN):

$$\frac{\text{TP}}{\text{TP} + \text{FN}}$$

Note that we divide the number of true positives by the number of individuals who actually have the disease tested for. The specificity (TNR) equals the number of true negatives (TN), divided by the sum of the number of true negatives (TN) and the number of false positives (FP):

$$\frac{\text{TN}}{\text{TN} + \text{FP}}$$

Note that we now divide the number of true negatives by the number of individuals who do not have the disease tested for. As an example, we observe from Table 3.2 that 126 individuals have a positive test result from the CT-scan and indeed have metastases in the liver. The number of true positives thus equals 126. The total number of individuals who actually have liver metastases equals $126 + 14 = 140$. The sensitivity of the scan now is computed to be $\frac{126}{140} = 0.90$; the specificity equals $\frac{133}{133+7} = 0.95$.

The sensitivity and specificity of a diagnostic test are sometimes combined into the *expected weight of evidence*, which basically is a measure of the overall usefulness of the test [20,38]. The expected weight of evidence $EW(d,T)$ of a test T with respect to the presence of the disease, is defined as

$$EW(d,T) = W(d,t) \cdot \Pr(t \mid d) + W(d,\overline{t}) \cdot \Pr(\overline{t} \mid d)$$

where t and $\overline{t}$ indicate the positive and negative result of the test T, respectively. $W(d,t)$ and $W(d,\overline{t})$ capture the weight of evidence for the value d of the disease variable D provided by the results t and $\overline{t}$, respectively, where $W(d,t)$ is defined as

$$W(d,t) = \log \frac{\Pr(t \mid d)}{\Pr(t \mid \overline{d})}$$

and $W(d,\overline{t})$ is defined analogously. Note that the weight of evidence $W(d,t)$ essentially is the logarithm of the likelihood ratio for a positive test result, which is the ratio of the probability of the test result t in the presence of the disease tested for and the probability of the test result t in the absence of the disease [13]:

$$LR(t) = \frac{\Pr(t \mid d)}{\Pr(t \mid \overline{d})}$$

Note that a low value for the likelihood ratio serves to indicate that the disease is excluded by a positive test result whereas higher values indicate that a positive result will tend to confirm the disease.

Table 3.1: The 2×2 probability table for the relationship between the absence or presence of metastases in a patient's liver and the result of a CT-scan of the liver.

		Metas-liver	
		yes/present	no/absent
CT-liver	yes/positive	$\Pr(t \mid d) = 0.90$	$\Pr(t \mid \bar{d}) = 0.05$
	no/negative	$\Pr(\bar{t} \mid d) = 0.10$	$\Pr(\bar{t} \mid \bar{d}) = 0.95$

Table 3.2: The 2×2 frequency table for the relationship between the absence or presence of metastases in a patient's liver and the results of a CT-scan of the liver.

		Metas-liver (*n*=280)	
		yes/present	no/absent
CT-liver	yes/positive	126 = TP	7 = FP
	no/negative	14 = FN	133 = TN

3.1.2 Predictive value

The sensitivity and specificity of a diagnostic test indicate how likely the test is to yield a result that matches the true presence or absence of the disease under study. Once a test result has become available, however, we are no longer interested in the probabilities of finding a positive or a negative result: we are much more interested in the effect of the test result found on the probability distribution over the disease variable. This effect is captured by the concept of *predictive value*. The standard concept of predictive value is once again defined in terms of a binary disease variable D and a binary test variable T. The *predictive value of a positive test result* (PVP) now is the probability $\Pr(d \mid t)$ of the disease indeed being present in a patient with a positive test result; it is defined as

$$\Pr(d \mid t) = \frac{\Pr(t \mid d) \cdot \Pr(d)}{\Pr(t \mid d) \cdot \Pr(d) + (1 - \Pr(\bar{t} \mid \bar{d})) \cdot (1 - \Pr(d))}$$

The predictive value of a positive test result can also be expressed in the numbers of true positives and false positives:

$$\text{PVP} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Note that, while in computing the sensitivity from a 2×2 frequency table we divided the number of true positives by the number of individuals who actually have the disease tested for, we now divide it by the number of individuals who have a positive test result. The *predictive value of a*

negative test result (PVN) is the probability $\Pr(\bar{d} \mid \bar{t})$ of the disease being absent in a negatively-tested patient; it is defined as

$$\Pr(\bar{d} \mid \bar{t}) = \frac{\Pr(\bar{t} \mid \bar{d}) \cdot \Pr(\bar{d})}{\Pr(\bar{t} \mid \bar{d}) \cdot \Pr(\bar{d}) + \Pr(\bar{t} \mid d) \cdot \Pr(d)}$$

Alternatively, this value can be expressed as

$$\text{PVN} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Note that we divide the number of true negatives by the number of individuals with a negative test result. As an example, we consider again the 2×2 frequency table shown in Table 3.2. From the table, we find that the number of true positives equals 126; the total number of individuals with a positive test result equals $126 + 7 = 133$. We compute the predictive value of a positive result from the CT-scan to be $\frac{126}{133} = 0.95$; the predictive value of a negative result equals $\frac{133}{133+14} = 0.95$.

3.1.3 Prevalence

In the previous sections we described the concepts of sensitivity, specificity and predictive value. A closely related concept is the *prevalence* of the disease. The prevalence is defined as the probability $\Pr(d)$ and equals the proportion of individuals in a group of individuals who possess the disease under study at a given point in time [11].

The sensitivity and specificity of a diagnostic test are not dependent upon the prevalence of the disease for which it has been designed. The two reliability characteristics therefore in essence are applicable to any population of patients. This property does not hold for the predictive values of a test's results. As an example, we consider again the CT-scan of a patient's upper abdomen to establish the presence or absence of secondary tumours in the liver. From Table 3.2, we found, in the previous section, the predictive value of a positive test result to be 0.95. This predictive value was computed from a population in which 140 of the total of 280 individuals have the disease; the prevalence of the disease thus equals 0.50. Now suppose that we have another population of individuals in which the disease under study is much more rare: $\Pr(d) = 0.09$. Table 3.3 records the findings for this population. Note that the sensitivity and specificity computed from the table are the same as before. The probability that an individual with a positive test result indeed has the disease tested for, has become smaller, however. The predictive value of a positive test result now equals

$$\Pr(d \mid t) = \frac{0.90 \cdot 0.09}{0.90 \cdot 0.09 + (1 - 0.95) \cdot (1 - 0.09)} \approx 0.64$$

thereby demonstrating the dependence of the predictive value of a test result upon the prevalence of the disease tested for.

Table 3.3: The 2×2 frequency table for the relationship between the absence or presence of metastases in the patient's liver and the results of a CT-scan of the liver, established from an alternative population.

		Metas-liver (n=1540)	
		yes/present	no/absent
CT-liver	yes/positive	126 = TP	70 = FP
	no/negative	14 = FN	1330 = TN

3.1.4 The costs of a diagnostic test

Upon selecting diagnostic tests for a patient, not only the characteristics of a test are important: also the potential risk for the patient of undergoing the test, the costs involved, and the time it takes to obtain a result from the test, for example, are of importance [18, 19]. While the financial costs and the time between ordering a test and obtaining the result are typically equal for every patient, the diagnostic value of the test's result and the risk involved are dependent upon the patient's specifics and upon the disease tested for [18, 19].

The term *costs* of a diagnostic test is generally used in a very broad sense, including numerous aspects. Often a distinction is made between direct costs and indirect costs [58]. The *direct costs* of a diagnostic test then are defined as the resources required to produce a result. These resources may include the costs associated with hospitalisation, administration, and equipment, as well as the routine costs of food, nursing care, supplies, operating rooms, anaesthesia, radiology, drugs, and physicians' professional fees. The *indirect costs* of a test involve all costs of illness other than direct costs. These indirect costs typically summarise the undesirable aspects of the test such as the risk involved, the discomfort of the test to the patient, and the possible consequences of the lag time between performing the test and obtaining its result. For the purpose of deciding which test to perform, both the direct and the indirects costs of a test are expressed numerically in utility units to allow for weighing and combining the diagnostic value of performing the test and the costs it involves [19]; we will return to this observation in Section 3.2.1.

3.2 The basics of test selection

A test-selection facility should be able to provide information about which tests are expected to provide the most valuable information in the present phase of the diagnostic process. The test-selection process should further result in a strategy that does not suffer from overtesting, that is, it should not select too many tests. Even more importantly, the resulting strategy should not halt the test-selection process too soon, since this may lead to misdiagnosis with possibly major consequences for both patients and physician. Automated test selection thus consists of three basic elements: a measure for establishing the usefulness of performing a particular test, a test-selection loop, and a criterion for deciding when to stop gathering further information. In

this section, we will review these three elements of automated test selection.

3.2.1 Test-selection measures

The usefulness of performing a diagnostic test is typically expressed using a *test-selection measure*. Such a measure can be based solely upon the expected decrease in diagnostic uncertainty after performing the test, or it can in addition incorporate the costs of the test. We start our review of test-selection measures with the former type.

In test selection, *information measures* are used for expressing diagnostic uncertainty. These measures in essence serve to assign a numerical value to the probability distribution over the main diagnostic variable. This numerical value attains its maximum when the diagnostic uncertainty is the highest, that is, for a uniform distribution over the diagnostic variable; it equals 0 when the diagnostic uncertainty is the lowest, that is, for any degenerate distribution over the diagnostic variable in which one of its values has been established with certainty. A commonly used information measure having these properties, is the Shannon entropy. More formally, we consider a disease variable D with the values $d_1, \ldots, d_m, m \geq 2$. The *Shannon entropy* H of the probability distribution $\Pr$ over D is defined as

$$H(\Pr(D)) = -\sum_{j=1,\ldots,m} \Pr(D = d_j) \cdot {}^2\log \Pr(D = d_j)$$

where $0 \cdot {}^2\log 0$ is taken to be 0. The Shannon entropy originates from information theory and describes the expected amount of information that is required to establish the value of the disease variable to certainty [56]. The amount of information that is required to establish a value d_j for D to certainty is expressed in bits and is weighted by the probability that d_j is the true value. A more general form of the Shannon entropy is Renyi's entropy [4]. *Renyi's entropy* of order α, with $\alpha > 0, \alpha \neq 1$, of the probability distribution over the disease variable D is defined as:

$$H_\alpha(\Pr(D)) = \frac{1}{1-\alpha} \cdot {}^2\log(\sum_{j=1,\ldots,m} \Pr(D = d_j)^\alpha)$$

For α approaching 1, Renyi's entropy approaches the Shannon entropy. We note that both probabilities close to 0 or 1 have a very small contribution to the Shannon entropy. For Renyi's entropy, however, we have that the larger the value for α, the higher the contribution of larger probabilities compared to that of smaller probabilities. In view of test selection, we have that for higher values for α, the test-selection process would show a tendency to halt too soon. Now, for test-selection purposes, the Shannon entropy of the probability distribution over the disease variable D is compared to the expected Shannon entropy of the posterior distribution over D after performing a test T. The test that is expected to result in the largest decrease in the Shannon entropy then is considered to be the best test to perform next [18, 19, 28].

An information measure that is closely related to the Shannon entropy is the Gini index [19]. It captures the difference between a degenerate probability distribution in which a value of the disease variable has been established to certainty and the current probability distribution over all

possible values. The Gini index exhibits the properties of an information measure mentioned above. When the diagnostic uncertainty is the highest, the value of the Gini index reaches its maximum. When the diagnostic uncertainty is the lowest, the value of the Gini index equals 0. More formally, the *Gini index* G of the probability distribution $\Pr$ over the disease variable D is defined as

$$G(\Pr(D)) = 1 - \sum_{j=1,\ldots,m} \Pr(D = d_j)^2$$

Note that by taking the various probabilities of a distribution to the power 2, the more extreme distributions effectively are assigned a smaller value than the more uniform distributions. For test-selection purposes again, the Gini index of the probability distribution over the disease variable D is compared to the expected Gini index of the posterior distribution over D after performing a test T. The test that is expected to result in the largest decrease in the Gini index then is considered to be the best test to perform next.

In addition to the Shannon entropy and the Gini index, sometimes the misclassification error is used for capturing uncertainty [24]. The *misclassification error* M of the probability distribution $\Pr$ over the disease variable D captures the difference between the probability of a certain diagnosis, that is, a probability equal to 1, and the probability of the most likely diagnosis. When the most likely value of the diagnostic variable has a large probability, the probability that this diagnosis will not be the correct one, is rather small. The value of the misclassification error then is small. On the other hand, when the probability of the most likely value of the diagnostic variable is relatively small, the probability of misclassifying is larger. The value of the misclassification error then is large. The misclassification error takes its maximum value when the probability of the most likely value is the smallest, that is, for a uniform distribution over the diagnostic variable. More formally, it is defined as

$$M(\Pr(D)) = 1 - \max\{\Pr(D = d_j) \mid j = 1, \ldots, m\}$$

For test-selection purposes, again, the misclassification error of the probability distribution over the disease variable D is compared to the expected misclassification error of the posterior distribution over D after performing a test T. The test that is expected to result in the largest decrease in the misclassification error then is considered to be the best test to perform next. The differences in behaviour of the Shannon entropy, the Gini index, and the misclassification error for test selection will be analysed in Chapter 5.

If the disease variable under study takes its value from among a finite set of numerical values, then sometimes the *variance* of the probabilities over these values is used as an information measure [28]. The variance can be described as the weighted average of the square of the distance of each value from the weighted mean of the variable. The measure then equals

$$V(\Pr(D = d_j)) = \sum_{j=1}^{m} (d_j - (\sum_{j=1}^{m} d_j \cdot \Pr(D = d_j)))^2 \cdot \Pr(D = d_j)$$

The more uniform the probability distribution over the diagnostic variable, the larger the variance; conversely, the closer the distribution is to degeneracy, the smaller the variance. For test-selection purposes, again, the variance of the probability distribution over the diagnostic variable D prior to performing a test and the expected variance of the distribution after performing a test are compared to derive information about which test is the best to perform next.

The various information measures reviewed above assign a numerical value to a probability distribution over the disease variable and compare two such distributions by comparing their numerical values. The *cross entropy*, or *Kullback-Leibler divergence*, allows for comparing two distributions more directly: it is a measure of the difference between two probability distributions [33]. The smaller the Kullback-Leibler divergence, the more similar the two distributions are, and vice versa. In the context of test selection, the Kullback-Leibler divergence is used for comparing the prior probability distribution $\Pr(D)$ over the disease variable D and the posterior distribution $\Pr(D \mid T_i = t_i^k)$ given a test result t_i^k for the test T_i. The Kullback-Leibler divergence then is defined as

$$KL(\Pr(D \mid T_i = t_i^k);\ \Pr(D)) = \sum_{j=1}^{m} \Pr(D = d_j \mid T_i = t_i^k) \cdot {}^2\log \frac{\Pr(D = d_j \mid T_i = t_i^k)}{\Pr(D = d_j)}$$

where the divergence is taken to be 0 whenever $\Pr(D = d_j) = 0$ since then also $\Pr(D = d_j \mid T_i = t_i^k) = 0$. Note that the Kullback-Leibler divergence is not symmetric and, hence, is not a measure of distance. Upon test selection, the Kullback-Leibler divergence serves to select the test that maximises the expected cross entropy between the probability distribution $\Pr(D)$ and the posterior distribution $\Pr(D \mid T_i)$ after performing test T_i.

So far, we have focused on information measures for test selection. An information measure in essence expresses just the uncertainty about the diagnosis and does not take any aspect of a test into consideration other than the decrease in diagnostic uncertainty it can occasion. Now, it has been argued by different researchers that a test should only be performed when the amount of information to be gained is worth the costs of performing the test [19, 38]. For test selection in a real-life decision-support system, the various information measures reviewed above can be readily extended to include, for example, the costs of testing, the discomfort for the patient, and the lag time between testing and obtaining the result. By subtracting the costs $C(T_i)$ of the test T_i from the expected decrease in diagnostic uncertainty, we establish the *expected profit* of performing the test [1, 29]. Glasziou and Hilden [19] introduce the concept of *net test worth* for diagnostic tests. The net test worth $W(T_i)$ of a diagnostic test T_i is defined as

$$W(T_i) = k \cdot I(\Pr(D \mid T_i)) - C(T_i)$$

where $I(\Pr(D \mid T_i))$ is the expected decrease in diagnostic uncertainty, measured for example using the Shannon entropy or the Gini index, and k is a scaling factor that represents the units of patient utility per unit of information. As an additional measure, they introduce the *ratio of net*

test worth to cost for a diagnostic test T_i:

$$\frac{W(T_i)}{C(T_i)} = \frac{k \cdot I(\Pr(D \mid T_i)) - C(T_i)}{C(T_i)} = k \cdot \frac{I(\Pr(D \mid T_i))}{C(T_i)} - 1$$

where k again is a scaling factor. Compared to the net test worth, this ratio fulfils the intuitive property that if the costs of an informative diagnostic test tend to zero, the test-selection measure tends to a universal maximum [19]. For test-selection purposes, the test that has the highest net worth or the highest ratio of worth to cost, alternatively, then is considered to be the best test to perform next.

The above test-selection measures in essence assume a constant ratio between the amount of information gained and the costs of a diagnostic test. It is possible, however, to weigh the gain in information and the costs of a test in a more complex way. To this end, every outcome of the patient's management, including possible misdiagnosis, complications, discomfort and financial costs, is given a numerical value, expressed in *utility units*. The worst possible outcome gets the utility 0 and the best possible outcome gets the utility 1. The utilities of the intermediate outcomes o_i are given by $u(o_i) = p_i$, where p_i is the probability for which the patient is indifferent between o_i for certain and a gamble with a probability p_i of obtaining the best outcome and a probability $1 - p_i$ of obtaining the worst outcome. For the disease variable D with values d_j, $j = 1, \ldots m$, the expected utility $EU(T_i)$ of performing a test T_i now equals [29, 58]:

$$EU(T_i) = \sum_{j=1,\ldots,m} EU(T_i, d_j) \cdot \Pr(d_j)$$

where $EU(T_i, d_j)$ captures the expected utility of performing the test if the patient actually has the disease d_j:

$$EU(T_i, d_j) = \sum_{k=1,\ldots,n} \sum_{r=1,\ldots,t} u(o_r) \cdot \Pr(o_r \mid T_i = t_i^k, d_j) \cdot \Pr(T_i = t_i^k \mid d_j)$$

Note that $EU(T_i, d_j)$ weighs the utility of every possible outcome o_r with the probability that this outcome occurs in the presence of the value d_j, given the various test results t_i^k of test T_i. The test that maximises the expected utility is now selected as the best test to perform next [1, 13, 19, 22, 25, 29].

We would like to note that although the use of utilities provides the most flexible way of establishing the usefulness of performing a specific test, it has a number of complicated consequences when employed for automated test selection. We observe that since utilities can differ among patients, the most informative test depends not only on the probability of the disease under study at a specific step in the test-selection process, but on the preferences of the patient as well. It is an open issue to which extent a patient is allowed to influence his management. Moreover, eliciting utilities is time-consuming and not trivial at all, especially not if it has to be done for each and every patient [32]. In this thesis, we have decided to focus on the use of an information measure such as the Shannon entropy and the Gini index, and to leave the use of utility-based test-selection measures for future research.

3.2.2 Myopic versus non-myopic test selection

Upon performing a full decision analysis, the decision maker reviews all possible sequences of diagnostic tests and their costs and benefits with respect to the patient's management and then chooses the strategy that maximises the expected utility [58, 66]. Performing such a full analysis however, is computationally very demanding. To render the selection of diagnostic tests computationally feasible, various different algorithms have been designed that approximate the analysis by a sequential approach in which one or more tests are selected at a time. We distinguish between two types of test-selection algorithm: *myopic* test selection, in which in each step a single test is selected, and *non-myopic* test selection, in which in each step multiple tests can be selected. We will begin our review of the various algorithms by focusing on the myopic ones.

Myopic algorithms
A myopic test-selection algorithm approximates a full decision analysis by selecting just the first test to be performed without taking the remainder of the steps in the patient's management into consideration; the best test to perform is computed using one of the test-selection measures described before. The selection of tests is re-iterated, thereby dynamically giving rise to a strategy like the full analysis. A myopic test-selection algorithm thus selects a single test T_i from a set of tests $\mathcal{T}$, prompts the user for its results, and selects the next test. In pseudo code, the myopic algorithm has the following general basic structure:

Myopic test selection

input:
$\mathcal{T}$: list of tests T_i

while $\mathcal{T} \neq \varnothing$ **do**
    compute most informative $T_i \in \mathcal{T}$
    prompt for evidence for T_i and process
    remove T_i from $\mathcal{T}$
od

Note that the basic myopic test-selection algorithm as presented above simply continues selecting diagnostic tests until the set $\mathcal{T}$ of all tests has been exhausted. For an automated test-selection facility, the algorithm has to be provided with a stopping criterion to decide upon when to stop gathering further information; we return to such stopping criteria in Section 3.2.3. Ben-Bassat [4] studied a myopic test-selection strategy for tests that are conditionally independent given the disease variable. He argues that, as long as information measures are used, a sequential test-selection algorithm will result in the same strategy as a full decision analysis whenever this property of conditional independence is satisfied.

While most automated test-selection facilities employ a myopic algorithm with a test-selection measure, a *critiquing facility* takes a somewhat different approach. The basic idea of critiquing is to focus test selection on a single disease value d_0. The goal now is to select diagnostic tests that maximise the probability of quickly accepting or declining d_0. To this end diagnostic tests are

selected that allow for discriminating d_0 from the other disease values. To measure the discriminative power of a diagnostic test, the expected weight of evidence is used. When the probability of the value d_0 becomes smaller than a prespecified threshold value, a new disease value will be focused upon. The critiquing procedure now selects a disease value, computes the expected weight of evidence for all tests that are still available, and continues the test-selection process until a hypothesis is accepted or no more tests are available [38]. We would like to note that, since the expected weight of evidence of a test is independent of the probability of the disease, the tests selected with respect to a specific disease value are the same for every patient. Only the choice of the disease values on which the process of test-selection is focused, will be specific for a patient. Critiquing thereby induces less flexibility in the selection of diagnostic tests than the decision-theoretic framework reviewed above.

Non-myopic algorithms

So far we addressed myopic test selection, in which tests are selected sequentially, on a one-by-one basis. Now, if the result of a selected test becomes available, it will cause a change in the expected benefit of all remaining tests. Unless costs are saved when performing multiple tests at once, therefore, performing more than one test at the same time will never be better than performing these tests sequentially [29]. From a medical perspective, almost always costs are saved by performing multiple tests: it can result in quicker diagnosis, for example, which can mean the difference between a curable and a non-curable cancer. When costs are taken into consideration, therefore, it is worthwhile to study the expected benefit of performing combinations of tests. Selecting the most informative combination of tests is known as non-myopic test selection. A non-myopic test-selection algorithm thus selects a single combination of tests C_i from the set of tests $\mathcal{T}$, prompts the user for its results and selects the next combination of tests. In pseudo code, the non-myopic algorithm has the following basic general structure:

Non-myopic test selection

input:
$\mathcal{T}$: list of tests

while $\mathcal{T} \neq \varnothing$ **do**
    compute most informative combination of tests $C_j \subseteq \mathcal{T}$
    prompt for evidence for all $T_i \in C_j$ and process
    remove all $T_i \in C_j$ from $\mathcal{T}$
od

For computing the most informative combination of tests, again one of the various test-selection measures presented in Section 3.2.1 is employed. The measure now has to address the probability distribution over the main diagnostic variable for multiple test results simultaneously.

As we will argue in Chapter 7, fully non-myopic test selection is much harder from a computational point of view than myopic test selection. To reduce the computational burden involved, various more restricted schemes of non-myopia in test selection have been studied. Heckerman,

for example, assumes in each step in the test-selection process that either no test or a predefined set of tests can be selected [25]. Jensen and Liang present an algorithm in which at first tests are selected myopically. As soon as all remaining tests have an expected profit smaller than 0, non-myopic test selection is performed, beginning with just all pairs of tests whose results are independent given the disease variable. If no pair of such tests has an expected profit larger than 0, then an approximate analysis of all remaining pairs of tests is performed under the assumption that these pairs are independent given the disease variable. Note that only two tests at a time are taken into consideration, thereby considerably reducing the computational burden involved in non-myopic test selection [29].

3.2.3 Stopping criteria

Automated test selection not just includes selecting appropriate tests in the different steps of a patient's management, it also involves deciding when to stop gathering further information. For this purpose, a test-selection facility is equipped with a stopping criterion. With this criterion, it decides after processing the evidence entered by the user, whether or not the diagnostic uncertainty has sufficiently decreased and no more tests should be ordered for the patient under consideration. A good stopping criterion can thus lead to fewer tests being performed, which results in lower financial costs, less discomfort for the patient, and perhaps even shorter waiting lists. Even more importantly, however, a good stopping criterion will prevent the test-selection algorithm to halt too early. Note that by stopping too early, the decision-support system may miss important information and reach an erroneous diagnosis, which may have far reaching consequences for the patient. When the costs of the various diagnostic tests are taken into consideration, these pose a more or less natural stopping criterion: the selection of diagnostic tests is continued as long as the expected benefits of a test outweigh its costs. In view of an information measure that does not take the tests' costs into consideration, no such natural criterion exists. Although many authors have described test-selection measures and test-selection algorithms, the issue of stopping criteria has not been addressed very often.

The most commonly used stopping criterion in view of an information measure considers, in each step of the test-selection process, the probability $\Pr(D = d_j)$ of the most likely value d_j of the disease variable D. If this probability is larger than a prespecified threshold value, the test-selection process is halted and no further evidence is gathered [21, 31]. As long as the probability is smaller than the threshold value, further tests are selected. The decision to stop gathering further information can also be based upon the difference between the probabilities of the two most likely values of the diagnostic variable. As soon as this difference becomes larger than a prespecified threshold value, the test-selection process is stopped [45]. Ben-Bassat [4] indicates that the test-selection process should halt when the current probability of error is less than a predefined threshold value. This probability of error is computed from a cost matrix for correct and incorrect decisions.

For all criteria, the threshold value used should be carefully determined. For example, a lower bound on the probability of the most likely value of the disease variable in case all test results are available, may be established to this end [59]. Similarly, an upper bound can be computed [39]. These bounds in essence are determined by computing the probability of the

most likely value of the diagnostic variable for all possible combinations of test results. As this can be a very expensive procedure from a computational point of view, McSherry proposes to determine apriori which test results do not have to be taken into account for the determination of the two bounds. He observes to this end that every test result is either a supporter or an opposer of the hypothesis. When $\Pr(t_i^k \mid d_j) \geq \Pr(t_i^k \mid d_l), j \neq l, l = 1, \ldots, m$, the test result t_i^k is a supporter of the disease value d_j; otherwise it is an opposer. A supporter of a disease value always increases the probability of that value, regardless of any other test results. For an opposer of the disease value, the contrary holds. He further shows that no combination of test results that minimises the probability of a specific diagnostic value, can contain a supporter of that value. Again, the contrary holds for opposers. Based upon these two consideration, fewer combinations of test results have to be taken into account upon computing the lower and upper bounds [39]. Although we support the idea of establishing such bounds for a real-life application, we have decided not to exploit the idea in our experiments, for computational reasons.

CHAPTER 4

Modelling reliability characteristics for probabilistic networks

Establishing a diagnosis for a patient in essence amounts to constructing a hypothesis about the disease the patient is suffering from. To this end, typically a number of diagnostic tests are performed. To establish the presence or absence of lung cancer, for example, an X-ray of the patient's chest is made. Diagnostic tests, however, generally do not unambiguously reveal the true condition of a patient. An X-ray, for example, can be difficult to interpret: a physician may easily overlook a small tumour and state a negative result, or state a positive result based upon a phantom image. To avoid misdiagnosis, therefore, the reliability characteristics of the various tests employed should be taken into consideration upon constructing a diagnostic hypothesis [17,58]. The specialised concepts of sensitivity and specificity have been designed for expressing the reliability of a test's results. The *sensitivity* of a diagnostic test is defined as the probability that a positive result is found in a patient who actually has the disease tested for; the test's *specificity* is the probability of finding a negative result in a patient without the disease. The medical literature lists these reliability characteristics for most diagnostic tests [8].

In the medical domain, decision-support systems are being developed for a wide range of diagnostic applications. These systems increasingly build upon a *probabilistic network* for capturing the domain's knowledge. A probabilistic network in essence is a representation of a joint probability distribution on a set of statistical variables, consisting of a graphical structure and an associated set of probabilities [28]. Since a probabilistic network uniquely defines a probability distribution, it allows for computing any probability of interest over its variables. Diagnostic reasoning with a probabilistic network now for example amounts to entering the symptoms, signs, and test results available for a patient into the network and computing the posterior probabilities for the various possible diseases. The disease with highest posterior probability then is taken for the diagnosis of the patient's condition.

For correct diagnostic reasoning, a probabilistic network has to capture the reliability characteristics with the various diagnostic tests that are represented in its graphical structure. In this chapter, we address the issue of modelling these characteristics. We argue that a probabilistic network has to distinguish explicitly between variables that represent test results on the one hand and variables that model the unobservable truth on the other hand. Failing to explicitly distinguish between these two types of variable would amount to assuming that test results unambiguously reflect the underlying truth and, hence, that diagnostic tests have a sensitivity and specificity both equal to 1.00 [64]. We further argue that the sensitivity and specificity of a test often have to be further detailed before they can be included in a probabilistic network: while the standard concepts are defined for binary variables, a network may include non-binary disease variables as well as non-binary test variables for which more detailed reliability characteristics are required. In addition, the network may require these characteristics to be stratified over several different subgroups of patients. We illustrate these as well as various closely related modelling issues by means of our real-life probabilistic network in the field of oesophageal cancer.

The chapter is organised as follows. Section 4.1 introduces the standard concepts of sensitivity and specificity, and describes how these concepts can be further detailed and stratified for inclusion in a probabilistic network. Section 4.2 discusses the closely related concept of predictive value of a test result, and describes how it can be computed from a probabilistic network. The chapter ends with our concluding observations in Section 4.3. This chapter is an based upon joint work with Linda C. van der Gaag [50, 51].

4.1 Test characteristics

In the medical domain, establishing a diagnosis for a patient typically involves performing a number of diagnostic tests. The results of these tests serve to reveal, to at least some extent, the underlying truth about the disease the patient is suffering from. As we already argued in the introduction, the reliability characteristics of these tests should be taken into account upon reasoning about the various possible diseases in order to avoid misdiagnosis. A probabilistic network developed for a diagnostic application, therefore, should explicitly capture these characteristics.

4.1.1 The standard concepts of sensitivity and specificity

The standard concepts of sensitivity and specificity have been designed to capture the uncertainty in the results of a diagnostic test, and thereby constitute the test's reliability characteristics. We recall from Chapter 3 that the concepts are defined in terms of two *binary* variables. We use the variable D to represent the presence, indicated by the value d, or absence, indicated by $\overline{d}$, of a specific disease. The variable T describes the result of an associated diagnostic test, where the value t models a *positive* test result, suggesting presence of the disease, and the value $\overline{t}$ models a *negative* result, suggesting absence of the disease. The *sensitivity* of a test now is defined as the probability that a positive test result is found in a patient who actually has the disease; more formally, it is defined as $\Pr(t \mid d)$. The complement $\Pr(\bar{t} \mid d)$ of the sensitivity is the probability that a negative result is found in a patient who has the disease tested for. The *specificity* of a test

is the probability $\Pr(\bar{t} \mid \overline{d})$ that the test yields a negative result for a patient without the disease. The complement $\Pr(t \mid \bar{d})$ of the specificity is the probability that a patient without the disease shows a positive test result. We note that a very sensitive test is good at detecting patients with the disease under consideration, while a very specific test is good at singling out patients who do not have the disease. The two concepts therefore pertain to different groups of patients and, hence, measure truly different reliability characteristics of a diagnostic test.

The medical literature lists the sensitivities and specificities for most diagnostic tests [8].

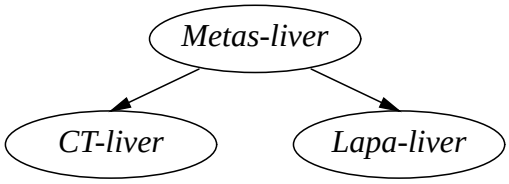
4.1.2 Sensitivity and specificity in a probabilistic network

To provide for modelling the reliability characteristics of diagnostic tests, a probabilistic network has to distinguish between variables that represent test results on the one hand and variables that represent the unobservable truth on the other hand. Only by such an explicit distinction can the relationship between a test's results and the underlying truth be represented. As an example of the two types of variable, Figure 4.1(a) depicts a fragment of the oesophageal cancer network. The variable *Metas-liver* describes the true presence or absence of secondary tumours in the liver of a patient; the variables *Lapa-liver* and *CT-liver* represent the results from a laparoscopic examination and from a CT-scan of the patient's liver, respectively. As the true presence or absence of liver metastases directly influences the results of the two diagnostic tests, the relationships between the variables are directed from *Metas-liver* to *Lapa-liver* and to *CT-liver* respectively. More formally, the represented relationships capture the assumption that the results of the two diagnostic tests are conditionally independent given the true presence or absence of liver metastases. We would like to note that this assumption of conditional independence is commonly made in medical decision making [66]; we will return to this observation in Section 4.1.6.

To conclude our example, we focus on the relationship between the variables *Metas-liver* and *CT-liver*. The strength of this relationship is captured by the four probabilities $p(\textit{CT-liver} = \text{yes} \mid \textit{Metas-liver} = \text{yes})$, $p(\textit{CT-liver} = \text{yes} \mid \textit{Metas-liver} = \text{no})$, $p(\textit{CT-liver} = \text{no} \mid \textit{Metas-liver} = \text{yes})$, and $p(\textit{CT-liver} = \text{no} \mid \textit{Metas-liver} = \text{no})$. We observe that the first and the last of these probabilities coincide with the sensitivity and specificity of the CT-scan; the other two probabilities are the complements of these characteristics. The table from Figure 4.1(b) now shows the probabilities that have been specified for the oesophageal cancer network: the CT-scan of the liver is stated to have a sensitivity of 0.90 and a specificity of 0.95. The results of the scan, therefore, are quite reliable.

4.1.3 More detailed sensitivity and specificity

The standard concepts of sensitivity and specificity are typically defined for binary variables only. The restriction to binary variables has its origin mainly in the use of these concepts. Physicians use the results from diagnostic tests, on seeing a patient, to decide whether or not to treat, or to decide whether or not to order another series of tests. These decisions in essence are binary and have to be taken within a relatively short time span. For taking such decisions, more detailed characteristics appear to be confusing rather than helpful [17]. There is no mathematical reason, however, for the restriction to binary variables.



(a)

		Metas-liver yes	no
CT-liver	yes	0.90	0.05
	no	0.10	0.95

(b)

Figure 4.1: A fragment of the oesophageal cancer network and some associated probabilities.

While practical for human decision making, the use of just binary variables may be an over-simplification of reality in view of computer-supported decision making. In a probabilistic network, for example, domain knowledge is typically represented in more detail, where statistical variables, whether modelling test results or the underlying truth, may have as many values as needed to accommodate the domain's intricacies. For modelling reliability characteristics that involve at least one non-binary variable, the standard concepts of sensitivity and specificity no longer suffice; the concepts have to be further detailed to provide for the more than two values. Because the sensitivity and specificity of a diagnostic test are well-defined in terms of probability theory, we can build upon this mathematical foundation for their refinement. We will begin by studying the reliability characteristics for a binary test variable and a non-binary disease variable. Then we address the situation where both variables are non-binary.

Detailing characteristics given a non-binary disease variable

We start by addressing a non-binary disease variable D with values $d_1, \ldots, d_n$, $n > 2$, and a binary test variable T with the values t and $\overline{t}$; for ease of exposition, we assume that the positive test result t pertains to a single disease value $d_i, 1 \leq i \leq n$. As an example, we consider from the oesophageal cancer network the variable *Invasion-organs*, that models the deepest point of invasion of the primary tumour into organs adjacent to the oesophagus. The variable can adopt one of the five values none, trachea, mediastinum, diaphragm and heart. In order to establish the depth of invasion of the primary tumour, typically a number of diagnostic tests are performed. The purpose of a bronchoscopy to this end, is to reveal invasion of the tumour into the trachea. The values of the test variable *Bronchoscopy* are yes, indicating that invasion into the trachea is visible, and no. The strength of the relationship between the two variables is captured by the conditional probabilities $p(\textit{Bronchoscopy} \mid \textit{Invasion-organs})$; the probabilities that have been specified for the oesophageal cancer network are shown in Table 4.1. We observe that the sensitivity of the bronchoscopy to tumour invasion into the trachea again coincides with one of the specified probabilities: the sensitivity equals $p(\textit{Bronchoscopy} = \text{yes} \mid \textit{Invasion-organs} = \text{trachea}) = 0.92$. The specificity $\Pr(\textit{Bronchoscopy} = \text{no} \mid \textit{Invasion-organs} \neq \text{trachea})$, however, is not captured

Table 4.1: The probabilities specified for the variable *Bronchoscopy*.

		Invasion-organs				
		none	trachea	mediastinum	diaphragm	heart
Bronchoscopy	yes	0.04	0.92	0.10	0.01	0.25
	no	0.96	0.08	0.90	0.99	0.75

by a single probability in the table. The four probabilities conditioning *Bronchoscopy* $=$ no on the values none, mediastinum, diaphragm and heart for the variable *Invasion-organs*, however, can be looked upon as more detailed specificities, indicating the probabilities that the test yields a negative result for patients in whom the tumour has invaded into the mediastinum, into the diaphragm, into the heart, and not into any organ beyond the oesophagus, respectively. From these more detailed specificities, however, the standard overall specificity can be reconstructed.

For a non-binary disease variable D with values $d_1, \ldots, d_n$, $n > 2$, and a binary test variable T with the values t, suggesting the presence of disease d_i, and $\overline{t}$, suggesting absence of d_i, we have for the overall specificity of the test that

$$\begin{aligned}\Pr(\overline{t} \mid \overline{d}_i) &= \Pr(\overline{t} \mid \textstyle\bigvee_{j=1,\ldots,n,j\neq i} d_j) = \\ &= \frac{\Pr(\overline{t} \wedge (\bigvee_{j=1,\ldots,n,j\neq i} d_j))}{\Pr(\bigvee_{j=1,\ldots,n,j\neq i} d_j)} = \frac{\Pr(\bigvee_{j=1,\ldots,n,j\neq i} (\overline{t} \wedge d_j))}{\Pr(\bigvee_{j=1,\ldots,n,j\neq i} d_j)}\end{aligned}$$

Since the values $d_j, j = 1, \ldots, n, j \neq i$, are mutually exclusive, we can build upon the additivity property of probability to arrive at

$$\Pr(\overline{t} \mid \overline{d}_i) \;=\; \frac{\sum_{j=1,\ldots,n,j\neq i} \Pr(\overline{t} \wedge d_j)}{\sum_{j=1,\ldots,n,j\neq i} \Pr(d_j)} = \frac{\sum_{j=1,\ldots,n,j\neq i} p(\overline{t} \mid d_j) \cdot \Pr(d_j)}{\sum_{j=1,\ldots,n,j\neq i} \Pr(d_j)}$$

Note that the probabilities $p(\overline{t} \mid d_j), j = 1, \ldots, n, j \neq i$, in the formula above coincide with probabilities specified for the test variable T. Also note that these probabilities convey the same type of information as the test's specificity be it in more detail, as we described above for the bronchoscopy. We further observe from the above formula that the more detailed specificities $p(\overline{t} \mid d_j)$ do not suffice for computing the overall specificity of the test: in addition, the prior probabilities $\Pr(d_j)$ for the disease variable D are required. Note that these prior probabilities are readily computed from the network. Since the overall specificity of a test now depends on the prior probability distribution on the disease variable and, hence, on the prevalences of the diseases under study, it is no longer generally applicable to any population of patients.

To return to our example, we recall that we are interested in the overall specificity of the bronchoscopy, that is, we are interested in the probability $\Pr($*Bronchoscopy* $=$ no $\mid$ *Invasion-organs* $\neq$ trachea$)$. From the oesophageal cancer network, we compute the prior probabilities of the various

Table 4.2: The probabilities specified for the variable *CT-organs*.

		Invasion-organs				
		none	trachea	mediastinum	diaphragm	heart
CT-organs	none	0.80	0.05	0.30	0.05	0.05
	trachea	0.05	0.58	0.10	0.10	0.10
	mediastinum	0.05	0.22	0.40	0.22	0.22
	diaphragm	0.05	0.05	0.10	0.53	0.10
	heart	0.05	0.10	0.10	0.10	0.53

possible depths of invasion to be

$$\begin{aligned}
&\Pr(\textit{Invasion-organs} = \text{none}) &&= 0.9193\\
&\Pr(\textit{Invasion-organs} = \text{trachea}) &&= 0.0299\\
&\Pr(\textit{Invasion-organs} = \text{mediastinum}) &&= 0.0234\\
&\Pr(\textit{Invasion-organs} = \text{diaphragm}) &&= 0.0202\\
&\Pr(\textit{Invasion-organs} = \text{heart}) &&= 0.0073
\end{aligned}$$

By substituting these probabilities in the above formula, we find

$$\frac{0.96 \cdot 0.9193 + 0.90 \cdot 0.0234 + 0.99 \cdot 0.0202 + 0.75 \cdot 0.0073}{0.9193 + 0.0234 + 0.0202 + 0.0073} \approx 0.96$$

for the overall specificity of the bronchoscopy.

We observe that this value approximates the probability $p(\textit{Bronchoscopy} = \text{no} \mid \textit{Invasion-organs} = \text{none})$. This observation is readily explained by considering the prior probabilities of the values of the variable *Invasion-organs*. These probabilities are quite small for an invasion of the primary tumour into the mediastinum, into the diaphragm, and into the heart. The prior probability of the absence of any invasion beyond the oesophageal wall, on the other hand, is very large. The influence of the probability $p(\textit{Bronchoscopy} = \text{no} \mid \textit{Invasion-organs} = \text{none})$ on the overall specificity therefore is much larger than the influences of the other more detailed specificities. Note that with another probability distribution on the values of the variable *Invasion-organs*, we would have found another overall specificity for the bronchoscopy.

Detailing characteristics for a non-binary disease variable

We now consider a non-binary disease variable D with values $d_1, \ldots, d_n, n > 2$, as before and a non-binary test variable T with values $t_1, \ldots, t_n, n > 2$; for ease of exposition, we assume that there is a one-to-one relation between a disease d_i and a test result t_i. An example from the oesophageal cancer network pertains to the disease variable *Invasion-organs*, that

models the deepest point of invasion of the primary tumour into organs adjacent to the oesophagus, and the test variable *CT-organs*, that models the deepest point of invasion of the tumour visible on a CT-scan of the patient's thorax. Both variables can adopt one of the five values none, trachea, mediastinum, diaphragm, and heart. The strength of the relationship between the two variables is captured by the conditional probabilities $p(\textit{CT-organs} \mid \textit{Invasion-organs})$; the probabilities that have been specified for the oesophageal cancer network are shown in Table 4.2. We observe that the five probabilities $p(\textit{CT-organs} = \text{trachea} \mid \textit{Invasion-organs} = \text{trachea})$, $p(\textit{CT-organs} = \text{mediastinum} \mid \textit{Invasion-organs} = \text{mediastinum})$, $p(\textit{CT-organs} = \text{diaphragm} \mid \textit{Invasion-organs} = \text{diaphragm})$, $p(\textit{CT-organs} = \text{heart} \mid \textit{Invasion-organs} = \text{heart})$, and $p(\textit{CT-organs} = \text{none} \mid \textit{Invasion-organs} = \text{none})$ can be looked upon as detailed sensitivities, indicating the probabilities that the CT-scan serves to identify the true depth of invasion of the primary tumour for patients in whom the tumour has invaded into the mediastinum, into the trachea, into the diaphragm, into the heart, and not into any organs beyond the oesophagus, respectively. No overall sensitivity is specified for the CT-scan of the thorax, however.

For a non-binary disease variable D with values $d_1, \ldots, d_n, n > 2$, and a non-binary test variable T with values $t_1, \ldots, t_n, n > 2$, the probabilities $p(t_i \mid d_i)$ are looked upon as the detailed sensitivities of the test to the various different diseases. We now define the overall sensitivity of the test to be the weighted sum

$$\sum_{i=1,\ldots,n} p(t_i \mid d_i) \cdot \Pr(d_i)$$

of these more detailed sensitivities. We note that for computing the overall sensitivity once again the prior probabilities $\Pr(d_i)$ for the disease variable are required. The overall sensitivity therefore does not apply to any population of patients. Using the above formula, we now compute the overall sensitivity of the CT-scan of the thorax to be

$$0.80 \cdot 0.919 + 0.58 \cdot 0.029 + 0.40 \cdot 0.023 + 0.53 \cdot 0.020 + 0.53 \cdot 0.007 \approx 0.78$$

While the detailed sensitivities $p(t_i \mid d_i)$ are directly available from the probabilities specified for the test variable T, the detailed specificities $p(\bar{t}_i \mid \bar{d}_i)$ are not. These detailed specificities are computed from

$$\Pr(\bar{t}_i \mid \bar{d}_i) = \frac{\sum_{j=1,\ldots,n,j\neq i} \sum_{k=1,\ldots,n,k\neq i} p(t_k \mid d_j) \cdot \Pr(d_j)}{\sum_{j=1,\ldots,n,j\neq i} \Pr(d_j)}$$

building upon similar observations as before. For the overall specificity of the test, we once again take the weighted sum

$$\sum_{i=1,\ldots,n} \Pr(\bar{t}_i \mid \bar{d}_i) \cdot \Pr(d_i)$$

of these detailed specificities. Note that, since a detailed specificity $\Pr(\bar{t}_i \mid \bar{d}_i)$ is defined with respect to a disease d_i, the various specificities are weighted by the prior probabilities of the diseases.

To return once again to our example, we note that we are interested in establishing the overall specificity of the CT-scan of a patient's thorax. The five detailed specificities of the CT-scan are computed to be

$$\begin{aligned}
&\Pr(\textit{CT-organs} \neq \text{none} \mid \textit{Invasion-organs} \neq \text{none}) && = 0.877\\
&\Pr(\textit{CT-organs} \neq \text{trachea} \mid \textit{Invasion-organs} \neq \text{trachea}) && = 0.947\\
&\Pr(\textit{CT-organs} \neq \text{mediastinum} \mid \textit{Invasion-organs} \neq \text{mediastinum}) && = 0.940\\
&\Pr(\textit{CT-organs} \neq \text{diaphragm} \mid \textit{Invasion-organs} \neq \text{diaphragm}) && = 0.948\\
&\Pr(\textit{CT-organs} \neq \text{heart} \mid \textit{Invasion-organs} \neq \text{heart}) && = 0.946
\end{aligned}$$

Using the prior probabilities of the various depths of invasion of the primary tumour, as given in the first part of this section, we now find the overall specificity of the CT-scan to be

$$0.88 \cdot 0.919 + 0.95 \cdot 0.029 + 0.94 \cdot 0.023 + 0.95 \cdot 0.020 + 0.95 \cdot 0.007 \approx 0.88$$

To conclude, we note that, since the values of both the disease variable and the test variable are mutually exclusive and collectively exhaustive, we have that the detailed sensitivities and specificities are readily defined for tests with a one-to-many or a many-to-one relation of their values to the values of the disease variable. By once again building upon their mathematical foundation, the overall reliability characteristics of such tests can be established from these detailed sensitivities and specificities.

4.1.4 Stratified sensitivity and specificity

In our domain of application, the reliability characteristics of some of the diagnostic tests depend on the patient's ability to swallow food. The results of a barium swallow with fluoroscopy to determine the presence or absence of a tracheoesophageal fistula, for example, are less reliable in a patient in whom the passage of food is seriously impaired. To further extend on the idea of including more detailed sensitivities and specificities in a probabilistic network, we now address modelling reliability characteristics that are dependent upon other variables than just the disease and test variables.

To capture the influence of another variable in addition to the disease variable, on a test's reliability characteristics, a probabilistic network has to explicitly capture the additional dependence in its graphical structure. In the oesophageal cancer network, for example, the variable *X-fistula*, modelling the result of a barium swallow, has an incoming arc not just from the disease variable *Fistula* but also from the variable *Passage*, as depicted in Figure 4.2. The relationship between the three variables is quantified by the conditional probabilities $p(\textit{X-fistula} \mid \textit{Fistula} \wedge \textit{Passage})$; the probabilities that have been specified for the oesophageal cancer network are shown in Table 4.3. We observe from the table that the sensitivity of the barium swallow does not coincide with one of the specified probabilities. While the probabilities $p(\textit{X-fistula} = \text{yes} \mid \textit{Fistula} =$

Table 4.3: The conditional probabilities specified for the variable *X-fistula*.

Fistula		yes				no			
Passage		solid	puree	liquid	none	solid	puree	liquid	none
	yes	0.88	0.88	0.88	0.85	0	0	0	0
X-fistula	no	0.10	0.10	0.10	0.05	1.00	1.00	1.00	0.90
	non-d.	0.02	0.02	0.02	0.10	0	0	0	0.10

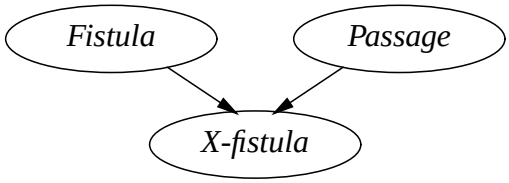


Figure 4.2: Another fragment of the oesophageal cancer network.

yes $\wedge$ *Passage*) for the various values of the variable *Passage* convey the same type of information as the standard sensitivity, they do so for different groups of patients. We say that the sensitivity is *stratified* over the variable *Passage*. The overall sensitivity and specificity of the barium swallow can now again be reconstructed from these stratified sensitivities and specificities.

We address a binary disease variable D and a binary test variable T. To indicate the different subgroups of patients, we define a new variable G with values $g_1, \ldots, g_m, m \geq 1$. For the variable T, the conditional probabilities $p(T \mid D \wedge G)$ are specified. By means of the conditioning rule of probability, we now find for the overall sensitivity of the test that

$$\Pr(t \mid d) = \sum_{i=1,\ldots,m} p(t \mid d \wedge g_i) \cdot \Pr(g_i \mid d)$$

Note that the probabilities $p(t \mid d \wedge g_i), i = 1, \ldots, m$, in the formula above coincide with probabilities specified for the test variable T. Also note that these probabilities convey the same type of information as the test's sensitivity be it stratified over different patient groups, as we described above for the barium swallow. We further observe that the overall sensitivity depends on the distribution of the patients over the various subgroups.

To return to our example, we recall that we are interested in the overall sensitivity of the barium swallow. For the stratified sensitivities, we observe that the barium swallow has a sensitivity of 0.88 for patients who are still able to swallow solid food; it equally has a sensitivity of 0.88 for patients who are only able to eat pureed food and for patients who can only swallow liquid

Table 4.4: The reliability characteristics of various tests computed from the oesophageal cancer network and found from the literature.

	sensitivity		specificity	
	experts	literature	experts	literature
Barium swallow	0.88	0.40 – 0.81	0.99	0.61 – 0.91
CT-lungs	0.90	0.80	0.95	0.80
CT-organs	0.78	0.40 – 0.81	0.88	0.61 – 0.91
Sonography neck	0.90	0.77	0.95	1.00
X-ray lungs	0.85	0.50 – 0.85	0.98	0.59 – 0.94

food; for patients who cannot swallow any food, the barium swallow has a sensitivity of 0.85. For the distribution of patients with a tracheoesophageal fistula over the four subgroups, we find

$$\begin{aligned}
\Pr(\textit{Passage} = \text{solid} \mid \textit{Fistula} = \text{yes}) &= 0.083 \\
\Pr(\textit{Passage} = \text{puree} \mid \textit{Fistula} = \text{yes}) &= 0.327 \\
\Pr(\textit{Passage} = \text{liquid} \mid \textit{Fistula} = \text{yes}) &= 0.493 \\
\Pr(\textit{Passage} = \text{none} \mid \textit{Fistula} = \text{yes}) &= 0.097
\end{aligned}$$

By substituting these probabilities in the above formula, we find the overall sensitivity of the barium swallow to be:

$$0.88 \cdot 0.083 + 0.88 \cdot 0.327 + 0.88 \cdot 0.493 + 0.85 \cdot 0.097 \approx 0.88$$

We consider again the binary disease variable D, the binary test variable T, and the variable G modelling the different patient groups under consideration. The probabilities $p(\overline{t} \mid \overline{d} \wedge g_i), i = 1, \ldots, m$, specified for T can be looked upon as stratified specificities, indicating the specificity of the test for the various groups of patients. By means of the conditioning rule of probability, we find for the overall specificity that

$$\Pr(\overline{t} \mid \overline{d}) = \sum_{i=1,\ldots,m} p(\overline{t} \mid \overline{d} \wedge g_i) \cdot \Pr(g_i \mid \overline{d})$$

Using the above formula, we compute the overall specificity of the barium swallow to be 0.99.

To conclude, we note that the concepts of stratified sensitivity and specificity can be extended to apply to non-binary disease and test variables as presented in the previous section.

4.1.5 Comparing characteristics against the literature

While the standard characteristics of diagnostic tests can typically be found in the literature, the more detailed and stratified sensitivities and specificities will not be readily available. These

characteristics will therefore have to be assessed by experts in the domain of application. The sensitivities and specificities found in the literature can then be used for comparison against the characteristics given by the domain experts. For some of the diagnostic tests that are commonly used for the staging of cancer of the oesophagus, Table 4.4 reports the overall sensitivities and specificities established from the oesophageal cancer network using the formulas presented in the previous sections; the table further reviews the sensitivities and specificities reported for these tests in the literature.

For a CT-scan of a patient's thorax to establish the depth of invasion of the primary tumour, for example, the literature reports a range of 0.40 to 0.81 for the sensitivity. This range is quite broad as some regions of the thorax are more clearly visible on the CT-scan than others; the CT-scan is more sensitive, for example, to metastases in the lungs than to invasion in the mediastinum. From our domain experts, we elicited the detailed sensitivities shown in Table 4.2. From this table, we observe that the sensitivity of the CT-scan has been assessed to range from 0.40 to 0.80. We note that this interval is consistent with the interval reported in the literature. The overall sensitivity of the test that we established to be equal to 0.78, therefore is also contained in the reported interval.

We would like to note that the reliability characteristics computed for the various diagnostic tests based upon the experts' assessments generally are higher than the sensitivities and specificities found in the literature for these tests [8]. From Table 4.4, we have for example that the literature reports for the CT-scan of the lungs both a sensitivity and a specificity of 0.80, while we found a sensitivity of 0.90 and a specificity of 0.95 from our experts. The literature notes that the reliability characteristics of a test tend to be highly dependent on a clinician's expertise and experience. Since for our domain of application the various characteristics were assessed by physicians from a highly specialised center for cancer treatment, the test results obtained may indeed be more reliable than in general, since tests are performed and interpreted by highly trained clinicians.

In general, experts' assessments of detailed and stratified sensitivities and specificities will not be as accurate as the reliability characteristics reported in the literature that have been established from clinical studies. We feel, however, that the possible inaccuracies are well compensated for by the additional diagnostic power that is gained by including non-binary variables in a probabilistic network.

4.1.6 A note on multiple tests

In medical practice, a diagnosis is often made on the basis of multiple tests. In the domain of oesophageal cancer, for example, to establish the presence or absence of secondary tumours in a patient's lungs, often a radiograph as well as a CT-scan of the thorax are made. The results of these tests are generally taken to be conditionally independent given the true condition of the patient. The results of multiple tests, however, may not be truly independent in practice. Although it is well known that assuming test results to be independent while in fact they are not, can lead to misdiagnosis, the assumption of independence is generally made for practical reasons. We recall from Section 4.1.2, where we discussed modelling the reliability characteristics of diagnostic tests, that for the oesophageal cancer network we equally assumed the results of multiple tests to

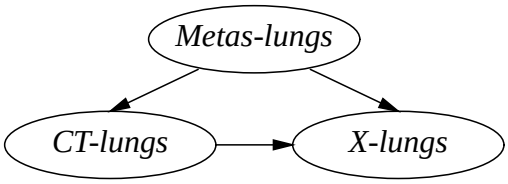


Figure 4.3: An alternative to a fragment of the oesophageal cancer network.

be conditionally independent.

We would like to note that a probabilistic network in essence provides for capturing the dependence between the results of two tests that cannot be considered independent. As an example, Figure 4.3 depicts an alternative to the fragment of the oesophageal cancer network pertaining to the different tests to establish the presence or absence of lungs metastases. The dependence between the results of the two tests is captured by an additional arc between the variables *CT-lungs* and *X-lungs*. We observe that for the variable *X-lungs* now the probabilities $p(\textit{X-lungs} \mid \textit{Metas-lungs} \wedge \textit{CT-lungs})$ have to be specified. Only rarely will these probabilities be known from clinical studies, however. The probabilities will moreover be difficult to assess by domain experts. As a consequence, in a probabilistic network, often conditional independence of test results is assumed as well.

4.2 Predictive value

In the previous sections, we studied the reliability characteristics of diagnostic tests, and focused on modelling these in a probabilistic network. The sensitivity and specificity of a diagnostic test indicate how likely the test is to yield a result that matches the true presence or absence of the disease under study. Once a test result has become available, however, we are no longer interested in the probabilities of finding a positive or a negative result: we are much more interested in the effects of the test result found on the probability distribution over the disease variable. This effect is captured by the concept of *predictive value*. The standard concept of predictive value is once again defined in terms of a binary disease variable D and a binary test variable T. The *predictive value of a positive test result* now is the probability $\Pr(d \mid t)$ of the disease indeed being present in a patient with a positive test result. The *predictive value of a negative test result* is the probability $\Pr(\bar{d} \mid \bar{t})$ of the disease being absent in a negatively-tested patient.

While the sensitivity and specificity of a diagnostic test are modelled explicitly in a probabilistic network, the predictive value of its results are not. The predictive value of a test's results, however, can be expressed in terms of the reliability characteristics of the test and the prior probability of the disease under consideration [58]. More specifically, the predictive value of a positive test result is defined by

$$\Pr(d \mid t) = \frac{p(t \mid d) \cdot \Pr(d)}{p(t \mid d) \cdot \Pr(d) + (1 - p(\bar{t} \mid \bar{d})) \cdot (1 - \Pr(d))}$$

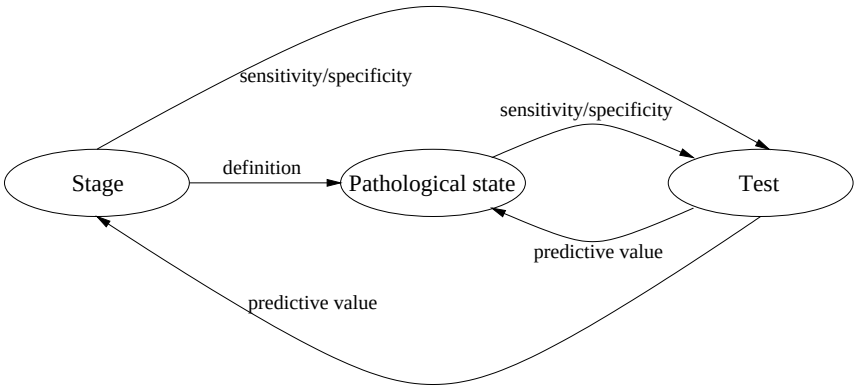


Figure 4.4: The relationships between the test, the true pathological state and the stage.

We observe that the conditional probabilities $p(t \mid d)$ and $p(\bar{t} \mid \bar{d})$ are the sensitivity and specificity of the test, which are specified for the variable T. The prior probability $\Pr(d)$ is readily computed from the network using an algorithm for probabilistic inference explained in Chapter 3. From Figure 4.1 and the observation that $\Pr(\textit{Metas-liver} = \text{yes}) = 0.21$, for example, we compute the predictive value of a positive test result of the CT-scan of the liver to be

$$\begin{aligned}\Pr(d \mid t) &= p(\textit{Metas-liver} = \text{yes} \mid \textit{CT-liver} = \text{yes}) = \\ &= \frac{0.90 \cdot 0.21}{0.90 \cdot 0.21 + (1 - 0.95) \cdot (1 - 0.21)} \approx 0.83\end{aligned}$$

Similar observations hold for the predictive value of a negative test result.

We would like to observe that we have modelled a test's reliability characteristics with respect to the underlying truth and not with respect to the main disease variable. For example, the reliability characteristics of a CT-scan of a patient's liver have been expressed in terms of the presence of secondary tumours in the liver, rather than in terms of the stage of the patient's cancer. The predictive value $p(\textit{Metas-liver} = \text{yes} \mid \textit{CT-liver} = \text{yes})$ of a positive result of the CT-scan then pertains to the presence of secondary tumours in the liver, while it is the probability distribution $\Pr(\textit{Stage} \mid \textit{CT-liver} = \text{yes})$ that is of interest. The predictive value thus pertains to the true pathological state and does not provide directly for establishing the main diagnosis. Since the various different tests are carefully embedded within the context of an overall network, however, the predictive value of a test result serves to indirectly update the probabilities of the main disease variable. Figure 4.4 captures the relationships between the concepts involved.

For ease of exposition, we address a binary main diagnostic variable S and a binary test variable T that provides for a binary variable D modelling the relevant pathological state. We are now interested in the main predictive value $\Pr(s \mid t)$. The probability $\Pr(s \mid t)$ equals

$$\Pr(s \mid t) = \frac{\Pr(t \mid s) \cdot \Pr(s)}{\Pr(t \mid s) \cdot \Pr(s) + (1 - \Pr(\bar{t} \mid \bar{s})) \cdot (1 - \Pr(s))}$$

The prior probability $\Pr(s)$ is readily computed from the network. For the probability $\Pr(t \mid s)$,

we observe that

$$\Pr(t \mid s) = \sum_{d_i \in \{d, \bar{d}\}} \Pr(t \mid d_i, s) \cdot \Pr(d_i \mid s)$$

Since the test result t is independent of s given d_i, we have that

$$\Pr(t \mid s) = \sum_{d_i \in \{d, \bar{d}\}} p(t \mid d_i) \cdot \Pr(d_i \mid s)$$

The probabilities $p(t \mid d)$ and $p(t \mid \bar{d})$ now are available from the characteristics that have been specified for the test T. The relationship between the presence or absence of the pathological state and the main diagnostic variable is modelled in the network, from which the probabilities $\Pr(d \mid s)$ and $\Pr(\bar{d} \mid s)$ are once again readily computed.

4.3 Conclusions

We addressed the issue of modelling reliability characteristics of diagnostic tests in probabilistic networks. In the medical domain, the reliability of a test is generally expressed in terms of its sensitivity and specificity. We argued that to capture these characteristics, a probabilistic network has to explicitly distinguish between variables that represent test results and variables that represent the underlying truth. Moreover, the standard concepts of sensitivity and specificity have to be further detailed to accommodate for non-binary disease and test variables. We also argued that stratified reliability characteristics are required for different groups of patients, when the result of a test does not just depend on the underlying truth but on a third variable as well. We showed that the standard concepts can be readily reconstructed from the more detailed, and possibly stratified, sensitivities and specificities.

Once a test result has become available, we are no longer interested in the test's sensitivity and specificity, but rather in the effects of the result on the probability distribution over the disease variable. We argued that this predictive value can be computed from the probabilistic network, using the detailed sensitivities and specificities. While in a probabilistic network predictive values pertain to the underlying truth and do not provide directly for establishing the main diagnosis, the context of the overall network provides for indirectly updating the probability distribution over the main disease variable with the test results entered.

CHAPTER 5

The behaviour of information measures for test selection

In many domains of application, sophisticated decision-support systems are being developed. Such systems are found for example in the medical domain where physicians have to establish a diagnosis and have to decide upon an appropriate therapy in relative uncertainty. To assist human decision makers in their complex reasoning processes, a diagnostic decision-support system is typically equipped with a test-selection facility that serves to indicate which tests had best be performed to decrease the uncertainty about the diagnosis [1, 4]. Most test-selection facilities select tests sequentially, that is, they select a single most informative test and await the user's input before selecting the next test to be performed.

The two most commonly used measures for capturing the uncertainty about the diagnosis in decision-support systems, are the Shannon entropy and the Gini index [19]; in other contexts, also the misclassification error is used for measuring uncertainty [24]. The three measures are defined for a probability distribution over a designated variable and express the expected amount of information that is required to establish the value of this variable with certainty. The Shannon entropy and the Gini index are generally considered to behave alike for test-selection purposes, in particular for variables with a small number of values [5]. In fact, common knowledge has it that the two measures are interchangeable in practice. In addition, the Shannon entropy and the Gini index are commonly acknowledged to be more sensitive to changes in the probability distribution over the main variable of interest than the misclassification error [6].

In this chapter, we compare the Shannon entropy, the Gini index and the misclassification error both from a fundamental and an experimental perspective. We show that the Gini index can be looked upon as an approximation of the Shannon entropy and that the misclassification error is an approximation of the Gini index. By studying the first derivatives of the three functions, moreover, we argue that for specific ranges of the probability distribution over the main

diagnostic variable, the Shannon entropy and the Gini index are indeed expected to behave alike. For the more extreme probability distributions, however, the two measures are expected to result in different test sequences. We furthermore argue that the misclassification error should not be used for test-selection purposes since it has a tendency to select tests randomly.

In addition to our fundamental analysis of the three measures, we studied the Shannon entropy and the Gini index from an experimental perspective. For this purpose, we implemented the two measures in a decision-support system for the domain of oesophageal cancer and performed test selection for 162 real patients. Upon analysing the sequences of tests yielded, we found that for 71% of the patients, already the first or second test selected differed between the two measures. In contrast with common knowledge, therefore, the Shannon entropy and the Gini index gave rise to rather different test-selection behaviour. All differences could be explained, however, from the insights that we had gained from our fundamental analysis of the Shannon entropy and the Gini index and their derivatives. By means of a sensitivity analysis, we further established the test-selection behaviour of the two information measures to be quite robust. We note that, although our experiment was conducted within the medical domain, the results of our fundamental analysis are domain-independent. A similar difference in behaviour is therefore expected to show in other domains as well.

The chapter is organised as follows. Section 5.1 reviews the Shannon entropy, the Gini index, and the misclassification error, and details how these measures are used for test selection. Section 5.2 presents our fundamental analysis of the three measures and of their first derivatives. In Section 5.3 we report on the results obtained with the Shannon entropy and the Gini index, and explain the observed differences. This chapter ends with our conclusions in Section 5.4. This chapter is based upon joint work with Linda C. van der Gaag [52].

5.1 Information measures and test selection

In a diagnostic decision-support system, test selection generally amounts to selecting, in a sequential manner, a test that is expected to yield the largest decrease in the uncertainty about the diagnosis. For capturing diagnostic uncertainty, typically an *information measure* is used. The three most commonly used measures are the Shannon entropy, the Gini index, and the misclassification error. These measures are defined for a probability distribution $\Pr$ over a set of statistical variables. We distinguish a diagnostic variable D, modelling the diagnoses of interest; the possible values of D are denoted d_j, $j = 1, \ldots, m, m \geq 2$. We further distinguish $n \geq 2$ test variables T_i, modelling diagnostic tests whose results can influence the uncertainty in D; the results of a test T_i are denoted t_i^k, $k = 1, \ldots, m_i$, $m_i \geq 2$. Each measure attains its maximum when the uncertainty about the value of the diagnostic variable is the largest, that is, when the probability distribution over this variable is a uniform distribution. For a distribution with $\Pr(d_i) = 1$ for some value d_i of D and $\Pr(d_j) = 0$ for all $d_i \neq d_j$, the uncertainty about the value of the diagnostic variable is resolved and the measures yield their minimum equal to 0.

The *Shannon entropy* H of the probability distribution over the diagnostic variable D is the expected amount of information that is required to establish the value of D with certainty; more

formally, the entropy is defined as

$$H(\Pr(D)) = -\sum_{j=1,\ldots,m} \Pr(D = d_j) \cdot {}^2\log \Pr(D = d_j)$$

where $0 \cdot {}^2\log 0$ is taken to be 0. For a binary diagnostic variable D modelling the presence or absence of a particular disease, with $p(D = \textit{present}) = 0.4$ and thus $p(D = \textit{absent}) = 0.6$, for example, the value of the Shannon entropy equals $-((0.40 \cdot {}^2\log 0.40) + (0.60 \cdot {}^2\log 0.60)) = -(0.40 \cdot -1.32 + 0.60 \cdot -0.74) = 0.97$.

Now suppose that some diagnostic test T_i is performed, and that the result t_i^k is yielded. Because of this additional information, the probability distribution over D will change from the prior distribution to the posterior distribution given $T_i = t_i^k$. The entropy of the distribution over D will then change as well, to the entropy of the posterior distribution:

$$H(\Pr(D \mid T_i = t_i^k)) = -\sum_{j=1,\ldots,m} \Pr(D = d_j \mid T_i = t_i^k) \cdot {}^2\log \Pr(D = d_j \mid T_i = t_i^k)$$

Returning to our example diagnostic variable D, we now suppose that the diagnostic test T_i has been performed and has yielded the result t_i^k which is construed as evidence of the disease being present. Further suppose that the probability $p(D = \textit{present} \mid T_i = t_i^k)$ of the disease being present given the test result t_i^k, equals 0.80. The entropy of the posterior distribution over D now equals $-((0.80 \cdot {}^2\log 0.80) + (0.20 \cdot {}^2\log 0.20)) = -(0.80 \cdot -0.32 + 0.20 \cdot -2.32) = 0.72$. The lower entropy value indicates that some of the uncertainty as to the presence or absence of the disease has been resolved by the evidence.

Prior to performing the test T_i, however, we do not know for certain what the result will be: each possible result t_i^k is yielded with a probability $\Pr(T_i = t_i^k)$. Before actually performing the test, therefore, we expect the entropy of the posterior probability distribution over D to be

$$H(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} H(\Pr(D \mid T_i = t_i^k)) \cdot \Pr(T_i = t_i^k)$$

We return again to our example variable D. We suppose that the diagnostic test T_i can result in one of the two values *positive* and *negative*. We further suppose that the probability that the test result is negative equals 0.3; the probability that the test result is positive thus equals 0.7. We now suppose that the entropy of the posterior distribution given a negative test result over the variable D equals $H(\Pr(D \mid T_i = \textit{negative})) = 0.94$; in the above we established $H(\Pr(D \mid T_i = \textit{positive}))$ to be 0.72. The expected entropy of the posterior probability distribution over D now equals $0.72 \cdot 0.7 + 0.94 \cdot 0.3 = 0.79$.

Given the definition of expected entropy, we now have that the decrease in uncertainty in the diagnostic variable D by performing the test T_i is expected to be $\widetilde{H}(T_i) = H(\Pr(D)) - H(\Pr(D \mid T_i))$. A test that maximises $\widetilde{H}(T_i)$ thus is the best test to perform. We assume that upon selecting a test that maximises the expected decrease in uncertainty, ties are broken at random.

The *Gini index* G of the probability distribution $\Pr$ over the diagnostic variable D is closely related to the Shannon entropy of the distribution; it is defined as

$$G(\Pr(D)) = 1 - \sum_{j=1,\ldots,m} \Pr(D = d_j)^2$$

The expected Gini index $G(\Pr(D \mid T_i))$ after performing a diagnostic test T_i is defined in the same way as the expected Shannon entropy $H(\Pr(D \mid T_i))$, that is, it is defined as the expected value of the Gini index where the expectation is taken over all possible test results:

$$G(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} G(\Pr(D \mid T_i = t_i^k)) \cdot \Pr(T_i = t_i^k)$$

The best test to perform once again is a test that is expected to result in the largest decrease of diagnostic uncertainty, that is, a test that maximises $\widetilde{G}(T_i) = G(\Pr(D)) - G(\Pr(D \mid T_i))$.

In addition to the Shannon entropy and the Gini index, sometimes the misclassification error is used for capturing uncertainty [24]; in the sequel we will argue that this measure is less suited for the purpose of test selection, however. The *misclassification error* M of the probability distribution $\Pr$ over the diagnostic variable D captures the difference between the probability of a certain diagnosis, that is, a probability equal to 1, and the probability of the most likely diagnosis; more formally, it is defined as

$$M(\Pr(D)) = 1 - \max\{\Pr(D = d_j) \mid j = 1, \ldots, m\}$$

The expected misclassification error after performing a diagnostic test T_i is

$$M(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} M(\Pr(D \mid T_i = t_i^k)) \cdot \Pr(T_i = t_i^k)$$

The decrease in diagnostic uncertainty in D by performing the test T_i thus is expected to be $\widetilde{M}(T_i) = M(\Pr(D)) - M(\Pr(D \mid T_i))$. A test that maximises $\widetilde{M}$ again is the best test to perform.

5.2 A fundamental analysis of the measures

To provide for predicting the test-selection behaviour of the Shannon entropy, the Gini index and the misclassification error, we study the three measures from a fundamental perspective. Since formally analysing a higher-dimensional function is not trivial, we begin by studying and comparing the behaviour of the measures for a binary diagnostic variable.

For a binary diagnostic variable D, with values d_1 and d_2, the Shannon entropy, the Gini index and the misclassification error can be written as

$$\begin{aligned} H(\Pr(D)) &= -\sum_{j=1,2} \Pr(D = d_j) \cdot {}^2\log \Pr(D = d_j) \\ &= -x \cdot {}^2\log x - (1 - x) \cdot {}^2\log(1 - x) \end{aligned}$$

$$G(\Pr(D)) = 1 - \sum_{j=1,2} \Pr(D = d_j)^2 = 2x - 2x^2$$

$$M(\Pr(D)) = 1 - \max\{\Pr(D = d_j \mid j = 1, 2\}$$

$$= \begin{cases} x & \text{, if } x \in [0, \frac{1}{2}] \\ 1 - x & \text{, if } x \in \langle \frac{1}{2}, 1] \end{cases}$$

where $x = \Pr(D = d_1)$. The three functions are depicted in the Figures 5.1(a), (b) and (c) respectively.

We begin our analysis by comparing the Shannon entropy and the Gini index. From Figure 5.1, we observe that the Shannon entropy has a higher value than the Gini index. To formally support this observation, we consider the second derivatives of the two functions:

$$H''(x) = -\frac{1}{x \cdot \ln 2} - \frac{1}{(1-x) \cdot \ln 2}$$

$$G''(x) = -4$$

We note that $H''(x) < G''(x)$ for all $0 < x < 1$. From this observation, we have that, in the interval $\langle 0, \frac{1}{2}\rangle$, the ascent of the Shannon entropy is steeper than that of the Gini index; in the interval $\langle \frac{1}{2}, 1\rangle$, the Shannon entropy shows a steeper descent than the Gini index. We further observe that the two functions attain the same value at $x = 0$ and at $x = 1$. We conclude that the two functions do not otherwise intersect and, hence, that $H(x) > G(x)$ for all $0 < x < 1$.

Upon further comparison of the Shannon entropy and the Gini index, we find that the Gini index can be regarded as an approximation of the Taylor expansion of the Shannon entropy. To formally support this observation, we build upon the Taylor expansion

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \ldots$$

of $\ln(1 + x)$ for $-1 < x \leq 1$. Using ${}^2\log x = \frac{\ln x}{\ln 2}$, we find that

$${}^2\log x = \frac{x-1}{\ln 2} - \frac{(x-1)^2}{2 \cdot \ln 2} + \frac{(x-1)^3}{3 \cdot \ln 2} - \frac{(x-1)^4}{4 \cdot \ln 2} + \ldots$$

for $0 < x \leq 1$, and

$${}^2\log(1 - x) = -\frac{x}{\ln 2} - \frac{x^2}{2 \cdot \ln 2} - \frac{x^3}{3 \cdot \ln 2} - \frac{x^4}{4 \cdot \ln 2} + \ldots$$

for $0 \leq x < 1$. Using the first term of these expansions, respectively, we find for the Shannon entropy that

$$H(x) \approx -x \cdot \frac{x-1}{\ln 2} - (1-x) \cdot \frac{-x}{\ln 2}$$

$$\approx 1.4 \cdot (2x - 2x^2)$$

For $0 < x < 1$, therefore, our Taylor approximation of the Shannon entropy differs from the Gini index by a multiplicative factor only. This property retains its validity if we take the first two terms of the Taylor expansions into account, yet with the multiplicative factor 1.8. We observe that, with $0 < x < 1$, the higher-order terms of the Taylor expansion have a rapidly decreasing value. We conclude that, for $0 < x < 1$, the Gini index may be considered an approximation of the Taylor expansion of the Shannon entropy by a multiplicative factor.

We now compare the misclassification error with the Gini index. Within the interval $[0, \frac{1}{2}]$, we have that

$$G(x) = 2x - 2x^2 \geq x$$

since from $0 \leq x \leq \frac{1}{2}$ we can conclude that $2x^2 \leq x$. The Gini index, therefore, lies above the misclassification error. Within the interval $\langle \frac{1}{2}, 1]$, we find that

$$G(x) = 2x - 2x^2 = 2x \cdot (1-x) > 1 - x$$

since from $\frac{1}{2} < x \leq 1$ we can conclude that $2x > 1$. Again, the Gini index lies above the misclassification error. We thus conclude that $G(x) \geq M(x)$. In fact, the misclassification error can be looked upon as a piece-wise linear interpolation of the three points $G(0), G(\frac{1}{2})$ and $G(1)$ of the Gini index.

Now, for test-selection purposes, we are not so much interested in the precise values that the Shannon entropy, the Gini index and the misclassification error attain for a specific probability distribution over the diagnostic variable D. We are more interested in the way they value a *shift* in the distribution that is occasioned by a test result. We therefore also study the first derivatives of the three functions:

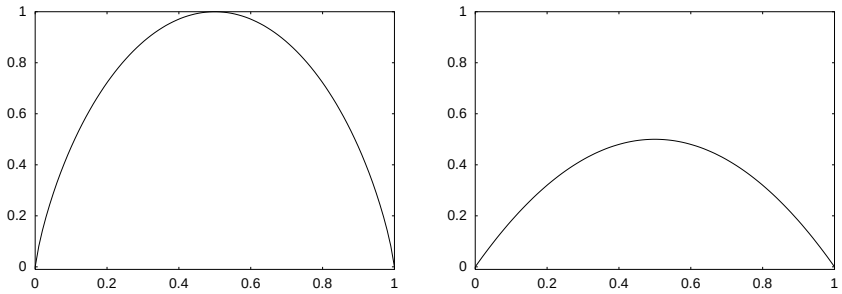
$$H'(x) = -{}^2\log x + {}^2\log(1-x)$$

$$G'(x) = 2 - 4x$$

$$M'(x) = \begin{cases} 1 & \text{, if } x \in [0, \frac{1}{2}\rangle \\ -1 & \text{, if } x \in \langle \frac{1}{2}, 1] \end{cases}$$

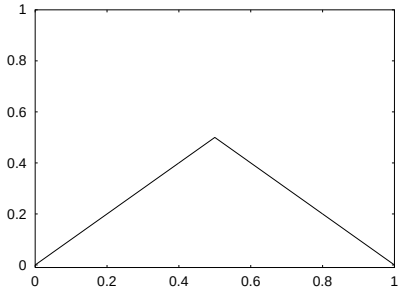
These derivative functions are depicted in the Figures 5.1(d), (e) and (f) respectively.

To compare the first derivatives of the Shannon entropy and the Gini index, we build, once again, on the Taylor expansion of ${}^2\log x$ for $0 < x \leq 1$, and of ${}^2\log(x-1)$ for $0 \leq x < 1$. Using

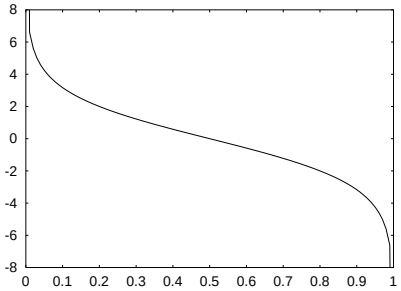


(a)

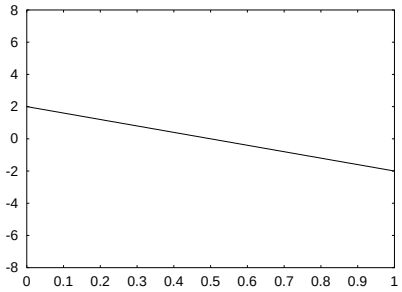
(b)



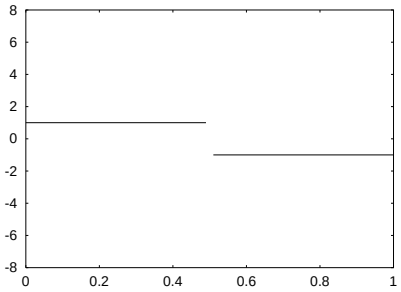
(c)



(d)



(e)



(f)

Figure 5.1: The Shannon entropy (a), the Gini index (b), and the misclassification error (c) of a distribution over a binary variable, and their first derivatives (d), (e) and (f).

the first term of these expansions respectively, we find for the first derivative of the Shannon entropy that

$$H'(x) \approx \frac{1-2x}{\ln 2} \approx 0.7 \cdot (2-4x)$$

We thus have that our Taylor approximation of the first derivative of the Shannon entropy differs from the first derivative of the Gini index by a multiplicative factor only. If we take the second term of the Taylor approximation into account, we find a multiplicative factor of 2.1.

From Figure 5.1(d), we observe that the first derivative of the Shannon entropy approximates a linear function in the middle part of the $[0, 1]$-interval, as suggested by the analysis above. For the more extreme values in the interval, however, this property no longer holds. Figure 5.2 illustrates this observation; the figure plots the function $H'(x)$ divided by $G'(x)$. To establish the boundaries of the interval within which the first derivative of the Shannon entropy is approximated by the first derivative of the Gini index, we consider once again the second derivatives of the two functions. We note that the maximum of the second derivative of the Shannon entropy equals -5.77 while the second derivative of the Gini index equals the constant function -4. Allowing a deviation of 10% off the value of the second derivative of the Gini index, we find that for values x for which $H''(x) \geq -6.17$, that is, for $x \in [0.37, 0.63]$, the first derivative of the Shannon entropy can be considered a linear function relative to the first derivative of the Gini index. We note that Figure 5.2 supports this observation: the figure suggests that within the established interval the two derivatives indeed differ by a constant multiplicative factor. For x approaching the extremes, however, this factor grows excessively in favour of $H'(x)$.

To compare the first derivatives of the Gini index and of the misclassification error, we begin by observing that G' is a linear function and M' is a piecewise constant function. We further observe that $G'(\frac{1}{4}) = M'(\frac{1}{4})$ and $G'(\frac{3}{4}) = M'(\frac{3}{4})$. We conclude that the first derivative of the misclassification error is a two-point approximation of the first derivative of the Gini index.

In view of the previous analysis, we briefly address the suitability of the misclassification error for the purpose of test selection. We observe that within the interval $x \in [0, \frac{1}{2}]$, we have that the misclassification error for the probability distribution over the diagnostic variable D equals $M(\Pr(D)) = \Pr(D = d_1) = x$. Now suppose that for a test variable T_i, we have that

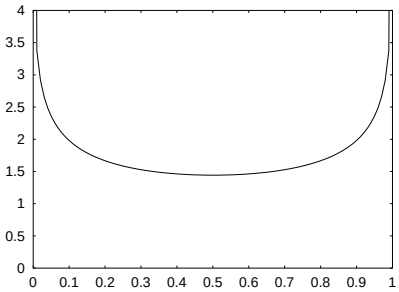


Figure 5.2: The first derivative $H'(x)$ of the Shannon entropy divided by the first derivative $G'(x)$ of the Gini index.

$\Pr(D = d_1 \mid T_i = t_i^k) \in [0, \frac{1}{2}]$ for all possible results t_i^k of T_i. We then find that the expected value of the misclassification error after performing the test equals

$$M(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} \Pr(D = d_1 \mid T_i = t_i^k) \cdot \Pr(T_i = t_i^k) = \Pr(D = d_1) = x$$

The expected misclassification error $M(\Pr(D \mid T_i))$ of the posterior distribution thus equals the misclassification error $M(\Pr(D))$ of the prior distribution, and the expected decrease in diagnostic uncertainty in D by performing the test T_i is $\widetilde{M}(T_i) = M(\Pr(D)) - M(\Pr(D \mid T_i)) = 0$. Similar observations hold for $\Pr(D = d_1) = x \in \langle \frac{1}{2}, 1]$. Only if the posterior probabilities of a diagnosis given the possible results t_i^k of T_i, are distributed over both intervals can the expected decrease in diagnostic uncertainty $\widetilde{M}(T_i)$ be larger than 0. Now, if at some stage in the test-selection process, for all remaining diagnostic tests the expected decrease in diagnostic uncertainty equals 0, the misclassification error will select a test at random. Since the probability distribution over the diagnostic variable is likely to become less uniform as the test-selection process progresses, the probability that a test will induce a shift to the other interval decreases. The misclassification error will then show a tendency to select tests rather arbitrarily; this tendency has been noted before by Breiman et al. [6]. We note that the tendency of the misclassification error to select tests at random may be quite undesirable for real-life decision-support systems.

We now review the implications of our findings for the test-selection behaviour of the Gini index and the Shannon entropy. The two measures value a test based upon the shifts that its results induce in the probability distribution over the diagnostic variable and upon the probabilities with which these results are expected to be found. Tests that induce a large shift in the probability distribution with a high probability, are valued as more informative than tests that result in a minimal shift with a high probability or in a large shift with just a small probability. Since the first derivative of the Gini index is a decreasing linear function, we find that it values a shift in distribution concavely by a constant factor. Since the first derivative of the Shannon entropy approximates a linear function between $x = 0.37$ and $x = 0.63$, it values a shift in a distribution where x stays between these two values in the same way as the Gini index. We conclude that the Shannon entropy and the Gini index will yield the same diagnostic test upon test selection as long as the tests under consideration are unlikely to result in a rather extreme distribution over the diagnostic variable. Since the Shannon entropy values a shift to an extreme distribution disproportionally more than the Gini index, the two measures may select different tests if a test is likely to result in such an extreme distribution. As an example, we consider a binary diagnostic variable modelling the presence or absence of a specific disease. Suppose that, at some stage in the test-selection process, we have that the probability of the disease being present is 0.38. The Shannon entropy of the distribution equals 0.96, while the value of the Gini index is 0.47. A positive test result of the diagnostic test T_1 shifts the probability of the disease being present to 0.02; a negative test result induces a shift of this probability to 0.47. The probability of a positive test result from test T_1 is 0.20; the probability of a negative test result thus is 0.80. The expected Shannon entropy of the posterior distribution over the diagnostic variable after performing test T_1 is 0.826; the expected Gini index of this distribution is 0.406. For the second diagnostic test T_2 we have that a positive test result shifts the probability of the disease being present to 0.22; a

negative test result from T_2 induces a shift of this probability to 0.62. The probability of finding a positive result from the test is 0.60; the probability of a negative result thus is 0.40. After performing test T_2, the Shannon entropy of the posterior distribution over the diagnostic variable is expected to be 0.839; the expected Gini index is 0.394. Based upon the expected decrease in uncertainty as to the presence or absence of the disease of interest, the Shannon entropy indicates test T_1 to be the most informative of the two tests, while the Gini index points to test T_2 as the more preferred one.

We note that several researchers [6, 19] have also described this difference in behaviour between the Gini index and the Shannon entropy. Glasziou and Hilden more specifically argue that the Shannon entropy overestimates the gain in information for shifts in an already extreme probability distribution.

So far, we studied and compared the Shannon entropy, the Gini index and the misclassification error in view of a binary diagnostic variable only. We now briefly focus on non-binary variables. For ease of reference, the three measures under study are depicted in Figure 5.3 for a three-valued diagnostic variable. Note that from the contour plots of the figure it is readily seen that again the Shannon entropy lies above the Gini index; the Gini index in turn lies above the misclassification error.

We consider a non-binary diagnostic variable D, with values $d_j, j = 1, \ldots, m, m \geq 2$. To study the test-selection behaviour of the three information measures, we are once again interested in the value they assign to a shift in the probability distribution over D. To gain insight in this value, we focus on a shift in a single probability $p_i = \Pr(D = d_i)$ under the assumption that the other probabilities $p_j = \Pr(D = d_j)$ retain their original proportions. We thus assume that each p_j is shifted to p'_j through

$$p'_j = p_j \cdot \frac{1 - x}{1 - p_i}$$

where x denotes the shifted probability of the diagnostic value d_i. Under this assumption, the Shannon entropy, the Gini index and the misclassification error respectively, can be written as

$$\begin{aligned} H(x) &= -\sum_{j=1,\ldots,m,j\neq i} p'_j \cdot {}^2\log p'_j - x \cdot {}^2\log x \\ &= -(1-x)\sum_{j=1,\ldots,m,j\neq i} \frac{p_j}{1-p_i} \cdot {}^2\log\left(\frac{p_j}{1-p_i}\cdot(1-x)\right) - x \cdot {}^2\log x \end{aligned}$$

$$\begin{aligned} G(x) &= 1 - x^2 - \sum_{j=1,\ldots,m,j\neq i} (p'_j)^2 = 1 - x^2 - \sum_{j=1,\ldots,m,j\neq i} \left(\frac{p_j}{1-p_i}\cdot(1-x)\right)^2 \\ &= (1-x^2) - (1-x)^2 \cdot \sum_{j=1,\ldots,m,j\neq i} \left(\frac{p_j}{1-p_i}\right)^2 \end{aligned}$$

$$\begin{aligned} M(x) &= 1 - \max\{x, p'_j \mid j = 1, \ldots, m, j \neq i\} \\ &= 1 - \max\{x, \left(\frac{p_j}{1-p_i}\right)\cdot(1-x) \mid j = 1, \ldots, m, j \neq i\} \end{aligned}$$

The (partial) derivatives with respect to x of the three measures are

$$H'(x) = -^2\log x - \frac{n}{\ln 2} + \sum_{j=1,\ldots,n,j\neq i}\left(\frac{p_j}{1-p_i}\cdot {}^2\log\left(\frac{p_j}{1-p_i}\cdot(1-x)\right)\right)$$

$$\begin{aligned} G'(x) =& -2x + 2\cdot(1-x)\cdot\sum_{j=1,\ldots,m,j\neq i}\left(\frac{p_j}{1-p_i}\right)^2 \\ =& -2\cdot\left(1+\left(\frac{p_j}{1-p_i}\right)^2\right)\cdot x + 2\cdot\sum_{j=1,\ldots,m,j\neq i}\left(\frac{p_j}{1-p_i}\right)^2 \end{aligned}$$

$$M'(x) = \begin{cases} 1 & \text{, if } x > \dfrac{p_j}{1-p_i}, j = 1,\ldots,m, j\neq i \\ \dfrac{p_j}{1-p_i} & \text{, if } p_j > p_k, k = 1,\ldots,m, k\neq j, j\neq i \end{cases}$$

We observe that the three derivatives are dependent of the original probability p_i. We further observe that, as for a binary diagnostic variable, the derivative of the Gini index is a linear function, while the derivative of the Shannon entropy is not. The derivative of the misclassification error, again, is a piecewise constant function. Under the assumption mentioned above, we may thus expect similar behaviour of the three measures for non-binary diagnostic variables as reviewed above for binary variables.

From the previous analysis, we have that, if the surfaces of Figure 5.3 are intersected with one of the planes $x = 0$, $y = 0$, or $x = 1 - y$, then the binary functions reviewed before are found. Intersecting the surfaces with a plane that is perpendicular to $x = 0$, $y = 0$, or $x = 1 - y$, yields a function that has the same shape as the binary function, yet is translated and restricted in its co-domain. Figure 5.4 shows, as an example, the function that results from intersecting the surface of the Gini index by the plane $x = 0.2$. Note that, with $x = 0.2$, the diagnostic uncertainty is maximal when the other two probabilities both are equal to 0.4, that is, when we have a uniform distribution of the remaining probability mass $1 - 0.2 = 0.8$. We further note that the diagnostic uncertainty can never be entirely resolved since there is some uncertainty left in the value d_i of the diagnostic variable. The minimum of the function consequently equals $1 - (0.8)^2 - (0.2)^2 = 0.32$, rather than 0.

5.3 The experiment

Building upon our fundamental analysis of the Shannon entropy, the Gini index and the misclassification error, we formulated, in the previous section, the differences to be expected in the test-selection behaviour of the three measures. Based upon our findings, we concluded that the misclassification error is not as suitable for test selection as the other two measures. In this section we therefore focus on the Shannon entropy and the Gini index. To study the differences between the two measures in a practical setting, we conducted a test-selection experiment using the measures in the context of a real-life decision-support system in oncology.

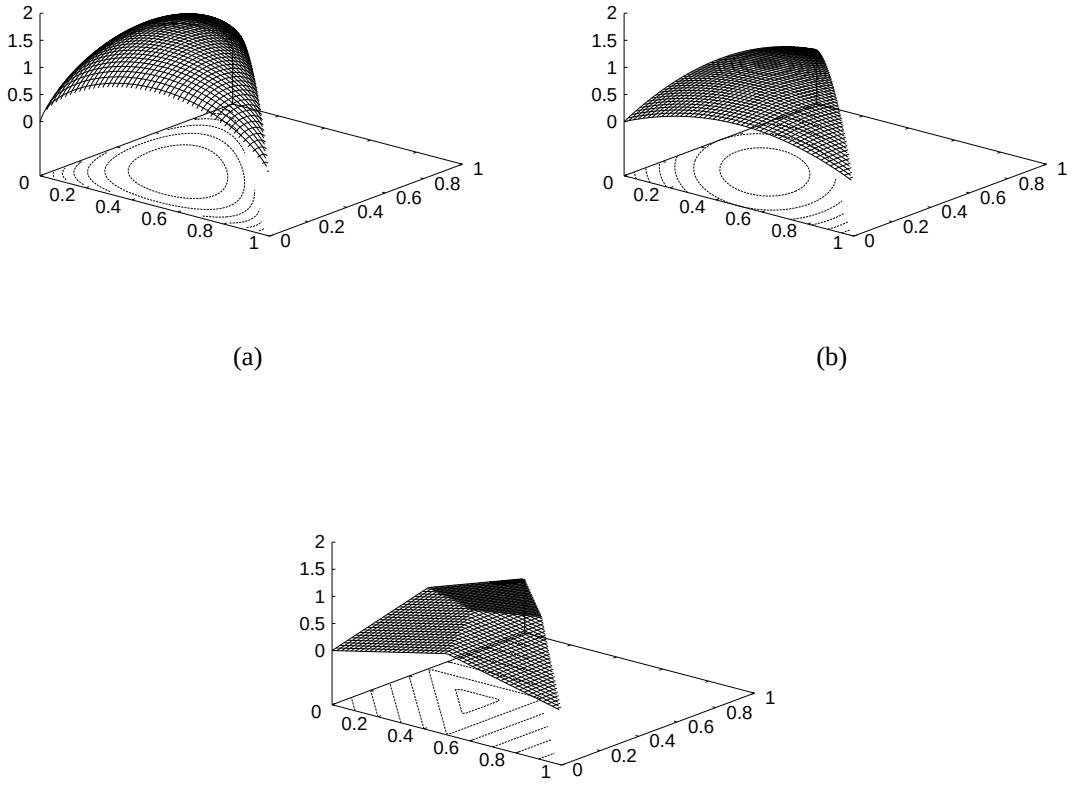


Figure 5.3: The Shannon entropy (a), the Gini index (b), and the misclassification error (c) for a three-valued variable.

5.3.1 The test-selection experiment

To study the test-selection behaviour of the Shannon entropy and the Gini index in the context of the oesophageal cancer network, we extended our decision-support system with a sequential selection facility. With the facility, we conducted two experiments. For the first experiment, we extended the oesophageal cancer network with a new binary variable *Operable* that summarises the six possible values of the original diagnostic variable *Stage* by classifying a patient's oesophageal cancer as operable or inoperable. For the second experiment, we used the test-selection facility in view of the original six-valued diagnostic variable. For our experiments, we had available the medical records of 162 patients diagnosed with cancer of the oesophagus. To simulate a realistic setting, we entered, for each patient, the results of a gastroscopic examination into the network prior to using the facility; in daily practice, the physicians also start selecting tests based upon the initial findings from this standard test.

An example
As an illustration of the results that we found from our first experiment, we discuss the test-selection behaviour of the two measures for a specific patient. When the test selection is started, the probability of the cancer of this patient being operable, equals 0.38; the Gini index of the distribution over the variable *Operable* equals 0.471 and the Shannon entropy equals 0.958. For the next test to perform, the Gini index and the Shannon entropy suggest different tests. The Shannon entropy indicates that a CT-scan of the liver is expected to result in the largest decrease in diagnostic uncertainty, whereas the Gini index selects an endosonography of the oesophageal wall. More specifically, the expected Shannon entropy is computed to be

$$\begin{aligned} &H(\Pr(\textit{Operable} \mid \textit{CT-liver=yes})) \cdot p(\textit{CT-liver=yes}) + \\ &H(\Pr(\textit{Operable} \mid \textit{CT-liver=no})) \cdot p(\textit{CT-liver=no}) = \\ &0.410 \cdot 0.231 + 0.998 \cdot 0.769 = 0.862 \end{aligned}$$

for the CT-scan and 0.899 for the endosonography; the expected values of the Gini index are 0.418 and 0.412 respectively.

To explain the observed difference in behaviour between the two measures, we study the shifts in the probability distribution over the diagnostic variable *Operable* that are occasioned by

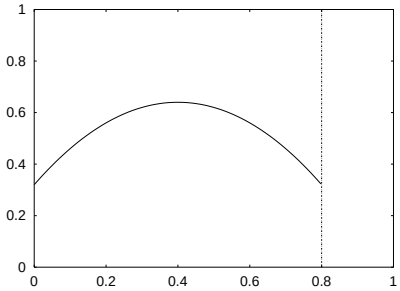
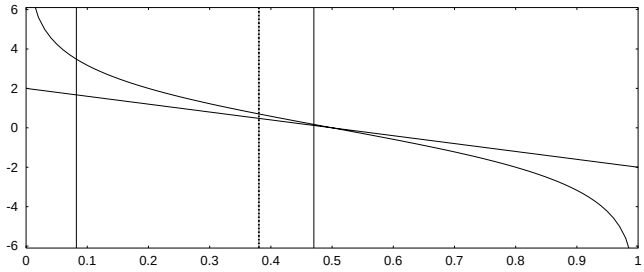
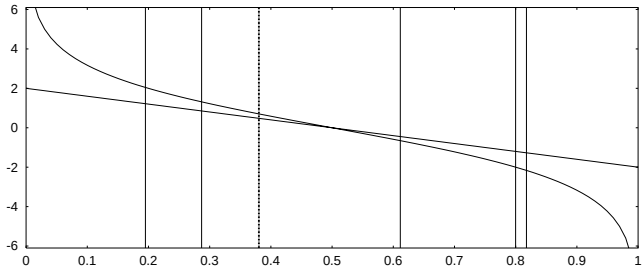


Figure 5.4: The Gini index for a three-valued variable, with $x = 0.2$.



(a)



(b)

Figure 5.5: The effects of the results of a CT-scan of the liver (a) and of an endosonography of the oesophageal wall (b), compared against the first derivatives of the Gini index and the Shannon entropy.

Table 5.1: The step, in the test-selection process, at which the Shannon entropy and the Gini index select different tests.

Step	*Frequency*	*Step*	*Frequency*	*Step*	*Frequency*
1	24	7	1	13	0
2	90	8	1	14	0
3	11	9	1	15	0
4	19	10	1	16	0
5	4	11	1	none	3
6	6	12	0		

the various test results. Figure 5.5(a) shows, by means of vertical lines, the shifts in distribution that are yielded by the two possible results of a CT-scan of the liver; the shifts occasioned by the five different values of the endosonography of the oesophageal wall are shown in Figure 5.5(b). The prior probability of the patient's cancer being operable is indicated by a bold vertical line in both figures. From Figure 5.5(a), we observe that the leftmost vertical line, indicating the probability 0.082 of the patient's tumour being operable given that the result of the CT-scan of the liver is *yes*, is well within the range in which the first derivative of the Shannon entropy no longer approximates a linear function. The shift in the probability distribution over the variable *Operable* that is occasioned by this test result, therefore, is valued much higher by the Shannon entropy than by the Gini index. The result moreover is relatively likely to be found, with a probability of 0.231. The result *no* of the CT-scan is valued more or less the same by both measures. Two results of the endosonography, on the other hand, are valued more or less concavely by a constant factor by the Shannon entropy as well as by the Gini index. The other three results of the endosonography are not within the range where they are valued more or less the same by the two measures. These three results have very low probabilities, of 0.034, 0.097 and 0.005, however. The result that serves to shift the probability of interest to 0.612, on the other hand, has a probability of 0.252, whereas the result that serves to yield a shift to 0.287 has a probability of 0.612. Note that although the probability 0.287 is not within $[0.37, 0.63]$, it is quite close to this interval. The shift to this probability is therefore valued more by the Shannon entropy than by the Gini index, yet not to a large extent. Since the shift occasioned by the endosonography is expected to result in a larger decrease of the uncertainty involved than that occasioned by the CT-scan of the liver, the Gini index selects the endosonography as the best test to perform. The expected decrease in diagnostic uncertainty by the CT-scan, however, is disproportionally larger with the Shannon entropy than with the Gini index, thereby explaining the Shannon entropy selecting the CT-scan. Note that these findings are conform the expectations from our fundamental analysis.

Overall results

So far, we discussed in detail the differences in test-selection behaviour of the two measures under study for a binary diagnostic variable. We also studied the differences in behaviour for the original six-valued diagnostic variable *Stage*. Table 5.1 summarises, over all patients, the

step in the test-selection process at which the Shannon entropy and the Gini index first selected a different diagnostic test. From the table we observe that for 24 patients (15%), already the first test differed. For 90 patients (56%), the measures selected the same diagnostic test for the first one to be performed, yet chose different tests for the second one. The range of tests selected in the first two steps was quite limited, however. The Shannon entropy selected the endosonography of the local region of the primary tumour for 44% of the patients as the most informative test, the endosonography of the oesophageal wall for 21% of the patients, and the CT-scan of the liver for 28% of the patients. The Gini index selected the endosonography of the local region of the primary tumour and of the oesophageal wall respectively, for 44% and 40% of the patients as the most informative test.

Since the Shannon entropy and the Gini index are commonly taken to be interchangeable for practical purposes, it is remarkable that for just three patients the two measures selected the same tests in the same order. The analysis from the previous section serves to explain why the two measures can select different tests. To explain the large number of differences found, we recall that, before the test-selection process is started for a patient, we entered the results from the gastroscopic examination into the network. Since these results tend not to influence the probability distribution over the diagnostic variable much, the test-selection process was started with a rather similar probability distribution for many patients. The example patient discussed in the previous section in fact belongs to this large group of similar patients.

5.3.2 Sensitivity analysis

To study the robustness of the differences in the test-selection behaviour of the Shannon entropy and the Gini index, we performed a sensitivity analysis of the oesophageal cancer network. In general, sensitivity analysis of a mathematical model amounts to investigating the effects of inaccuracies in the model's parameters by systematically varying the values of these parameters and studying the impact of the variation on the model's output [41]. In view of test selection more specifically, the effect of parameter variation on the selected tests is studied.

The oesophageal cancer network includes some 1000 parameters, 280 of which capture the detailed and possibly stratified sensitivities and specificities of the modelled diagnostic tests [50]. In our sensitivity analysis, we varied each of these 280 parameters, from 0 to 1. Using the two information measures, we selected the most informative test with respect to the binary diagnostic variable as described before, and identified for each parameter for which value the most informative test changed, building upon the concept of admissible deviation [60]. For 75 of the 280 parameters (27%), we found that upon variation at least one of the information measures suggested another test than with the original value of the parameter. For example, for the parameter $x = \Pr(\textit{CT-liver=no} \mid \textit{Metas-liver=yes})$, with an original value of 0.10, the Shannon entropy selected the CT-scan of the liver as the most informative test for $x < 0.2$; for $x \geq 0.2$, it selected the endosonography of the oesophageal wall. The tests selected by the information measures were not sensitive to variation of the other 205 parameters. We note that only if the original value of a parameter was within the leftmost or rightmost quarter of the $[0, 1]$-interval, could varying the parameter to a larger or smaller value result in selecting different diagnostic tests. This finding corroborates the theoretical properties of sensitivity functions established by

Renooij and Van der Gaag [47]; they show more specifically that variation of any parameter with an original value in the mid-half of the $[0, 1]$-interval can have only a limited effect on an output probability of interest.

We further noticed that the parameters whose variation could change the selected test, corresponded with the tests that were selected in the first three steps of the regular test-selection experiment. This finding can easily be explained. A highly informative test is likely to induce a large shift in the probability distribution over the diagnostic variable, with a relatively high probability. If the sensitivity and specificity characteristics of such a test are varied, the shift in the probability distribution over the diagnostic variable may become considerably smaller, thereby effectively decreasing the test's informativeness. An information measure may then select another test as the most informative one. The only exception to the previous observation pertains to the laparoscopy of the liver. Although this test was never selected within the first three steps of the test-selection process, the sensitivity analysis revealed that by varying its characteristics, the test could become highly informative. The laparoscopy, like the CT-scan, provides information about the absence or presence of metastasis in the liver, which in turn is highly indicative of the stage of the cancer. The characteristics of the two tests differ considerably, however. While the CT-scan of the liver is a highly reliable test, the laparoscopy is not. When the parameters for the laparoscopy of the liver are varied to the extent to which these approximate the parameters for the CT-scan, it is readily explained why the selection of the test is sensitive to the variation of its parameter values. A similar observation does not hold for the tests that serve to give insight in the absence or presence of metastases in a patient's lungs, since these have been assessed to have a lower predictive value on the stage of the patient's oesophageal cancer.

From the results of our sensitivity analysis, we conclude that the test-selection process based upon the two information measures is fairly robust. Only when the sensitivity and specificity characteristics of highly informative tests are varied quite drastically, can the measures be induced to select a different test. This observation suggests that the difference in behaviour between the two measures can indeed be attributed to their mathematical properties and does not originate from the specific parameter values in our network.

5.4 Conclusions

Diagnostic decision-support systems are often equipped with a test-selection facility that builds upon an information measure for assessing diagnostic uncertainty. In this chapter we studied the Shannon entropy, the Gini index and the misclassification error for this purpose. We analysed the three measures from a fundamental perspective and investigated their first derivatives. We argued that the Gini index can be looked upon as an approximation of the Shannon entropy and that the misclassification error is a two-point approximation of the Gini index.

With respect to their test-selection behaviour, we observed that, although a shift in many probability distributions over the diagnostic variable is valued similarly by the Gini index and the Shannon entropy, a shift to rather extreme distributions is valued much higher by the Shannon entropy than by the Gini index. Based upon this observation, the two measures are expected, at least occasionally, to select different tests. We feel that, despite their possible difference in behaviour, both measures are equally suited for use in a decision-support system. The charac-

teristics of the application domain under study then determine which of the two measures had best be used. We furthermore concluded that the misclassification error should not be used for test-selection purposes due to its tendency to select tests randomly when all possible shifts in the probability distribution over the diagnostic variable are within the same half of the $[0, 1]$-interval.

We conducted an experiment to study the behaviour of the Shannon entropy and the Gini index in a real-life setting. The results from our experiment served to corroborate the differences in behaviour expected from our more fundamental analysis. A sensitivity analysis with respect to the selection of tests based upon the two information measures, moreover, showed that test selection based on the Shannon entropy and the Gini index is quite robust. Since our fundamental analysis and expectations of the two information measures are independent of our domain of application, we feel that the differences in test-selection behaviour observed in our experiments in the domain of oesophageal cancer, are likely to show in other domains as well.

CHAPTER 6

Eliciting test-selection strategies

Decision-support systems are being developed for a wide range of domains. To support the reasoning processes in its domain of application, such a system often includes a facility for selecting tests. In the medical domain, a decision-support system may, for example, suggest a sequence of diagnostic tests to be performed in order to reduce the uncertainty about a patient's true condition. The test-selection strategy thereby provides support for the task of diagnostic reasoning. In most decision-support systems, a facility is offered in which tests are suggested sequentially, that is, on a one-by-one basis. The system then suggests a single test to be performed and awaits the user's input; after taking the test's result into account, the system suggests a subsequent test, and so on.

With the help of two experts in gastrointestinal oncology a decision-support system for the domain of oesophageal cancer has been developed during the last decade [61]; the oesophageal cancer network embedded in this system has been described in detail in Chapter 2. The oesophageal cancer system at present does not support the selection of diagnostic tests. For a specific patient, the attending physician orders a number of tests, based upon his or her own judgement, and simply enters the results into the system; the system does not provide the physician with information about which tests would be relevant and should be considered next for the patient under consideration. Building upon the system's mathematical foundation, however, a sequential test-selection strategy could easily be designed. Such a strategy would select, on a one-by-one basis, the test that is the most informative, for example in terms of entropy reduction, given the already available patient specifics [1, 13]; we elaborated on such a sequential strategy in Chapter 5. Upon working with the system, however, we noticed that in daily practice diagnostic tests are not selected on a one-by-one basis but are ordered in packages instead. We concluded that a sequential test-selection strategy would be an oversimplification of our experts' problem-solving practice.

With the advance of protocols for many fields of medicine, guidelines are becoming avail-

able that prescribe, to at least some extent, the diagnostic tests that physicians should perform at various intervals in their reasoning processes. Unfortunately, a detailed protocol that could constitute the basis for a test-selection facility for our decision-support system for oesophageal cancer, is not available as yet. We decided to design a test-selection facility for our system that would build upon the arguments used by the experts for deciding whether or not to order specific tests. The resulting facility would thus more closely fit in with the strategies for test selection currently used in the domain than a standard sequential test-selection facility.

To acquire knowledge about the actual test-selection strategy employed by our experts and about the arguments underlying their strategy more specifically, we used an elicitation method that combined several different techniques for knowledge elicitation [15]. The method consisted of three main interviews. The first of these was an unstructured interview that was aimed at providing insight in the overall strategy used by the experts. The second interview was a structured interview in which further details were acquired. In this latter interview, the experts' problem-solving practice was carefully simulated by means of cards, or vignettes, describing realistic patient cases. By simulating daily routine, we aimed to exclude, as much as possible, the various different biases that could possibly originate from the elicitation method used. We note that the idea of following up an unstructured interview by a structured one has been proposed before, for example in Cognitive Task Analysis [48]. The third and last interview served to review and refine the flowchart that had resulted from the second interview. We used the elicitation method with the two experts in our domain of application. We found that the method, and the use of carefully designed patient cases more specifically, closely fitted in with the experts' daily practice.

Since we will be using the arguments underlying our experts' problem-solving strategy to build a facility for automated test selection, we had to ascertain that we had elicited their strategy at a sufficient level of detail. To this end, we studied the medical records of real patients and compared these against the flowchart that we had constructed after the third interview. We found that the results of the comparison were influenced by various issues. For example, changes in medical practice over time and patient data originating from different hospitals are likely to have obscured the results to a considerable extent. Missing values, moreover, had seriously hampered the comparison. As a consequence, we were not able to derive any definite conclusions from the comparison.

Since a test-selection facility offered by a decision-support system in general should support physicians in their daily problem-solving practice, we feel that for the design of such a facility, knowledge about the actual strategies employed in the domain of application should be elicited from experts; a standard, sequential strategy may then turn out to deviate too much from daily routines to be acceptable. Our experiences in the domain of oesophageal cancer have demonstrated that the knowledge required for the design of a tailored test-selection facility can be feasibly acquired: with our elicitation method, we were able to elicit the arguments underlying our experts' strategy in little time. We would like to note that the issues that hampered the evaluation of the flowchart against the data, are not reserved for our field of application. We feel that these issues are likely to arise in many retrospective data collections and should be carefully analysed.

The chapter is organised as follows. In Section 6.1, we provide some preliminaries on oesophageal cancer and its therapies. In Section 6.2, we give an overview of the method that we used for eliciting our experts' test-selection strategy. In Section 6.3, we describe the results that

we obtained from the first, unstructured interview. Section 6.4 reports on the second, structured interview. Section 6.5 reports on the third and last interview that we had with the domain experts. In Section 6.6 we discuss the results from comparing the elicited flowchart to the data we had available. The chapter ends with our concluding observations in Section 6.7. This chapter is based upon joint work with Linda C. van der Gaag, Cilia L.M. Witteman, Berthe M.P. Aleman and Babs G. Taal [53, 54].

6.1 Preliminaries

As described in Chapter 2, generally a number of diagnostic tests are performed to establish the stage of a patient's oesophageal cancer. Various different tests are available, giving insight in different aspects of the cancer. A gastroscopic examination, for example, provides information about the presentation characteristics of the primary tumour, which include its length and its location in the oesophagus. A biopsy reveals the histological, or cell, type of the tumour. A laparoscopic examination of the liver, a CT-scan of the liver and of the lungs, as well as an X-ray of the lungs provide evidence about the presence or absence of haematogenous metastases. An endosonographic examination serves to yield information about the depth of invasion of the primary tumour into the oesophageal wall. The available tests differ considerably with respect to their reliability characteristics. The sensitivity and specificity characteristics of a diagnostic test play an important role in the selection of tests in normative decision making, since these characteristics indicate how useful, or how informative, a negative or a positive result of the test actually is [58].

For patients suffering from oesophageal cancer, various different treatment alternatives are available. These alternatives include surgical removal of the primary tumour, administering radiotherapy, and positioning a prosthesis. Providing a therapy aims at removal or reduction of the patient's primary tumour to prolong life expectancy and to improve the passage of food through the oesophagus. The therapies differ in the extent to which these effects can be attained, however. The main goal of a surgical procedure is to attain a better life expectancy for a patient, that is, the procedure is curative in nature. Positioning a prosthesis in the oesophagus, on the other hand, is a palliative procedure that cannot improve life expectancy: it is performed merely to relieve the patient's swallowing problems. Radiotherapy can be administered in a curative regime, aimed at prolonging a patient's life, as well as in a palliative regime, aimed just at improving the patient's quality of life. The most preferred treatment in essence is to provide a curative therapy; of these curative therapies, a surgical removal of the primary tumour is preferred to a curative regime of radiotherapy. Providing a therapy, however, is often accompanied not just by beneficial effects but also by complications. These complications can be quite serious and may even prove to be fatal. The beneficial effects and complications to be expected from the different therapies for a specific patient depend on the general condition of the patient, on the characteristics of his or her primary tumour, on the depth of invasion of the tumour into the oesophageal wall and neighbouring organs, and on the extent of metastasis of the cancer. If serious complications are expected for the patient, the attending physician may decide to abstain from providing a curative therapy and to administer one of the palliative treatment alternatives.

6.2 A method for eliciting test-selection strategies

Before describing the method that we used for eliciting test-selection strategies and before introducing the setting in which we used it, we briefly review some well-known elicitation methods.

The background
For knowledge acquisition, generally a distinction is made between methods that are aimed at eliciting object knowledge (knowing that) and methods with which to elicit problem-solving knowledge (knowing how). For finding out how our domain experts select diagnostic tests, we needed a method that focused on the latter type of knowledge. Several such methods are available, among which observation methods, methods for eliciting verbal reports, and think-aloud methods are the most well known [49, 65].

Observation amounts to recording the behaviour that is exhibited by experts while they are solving problem situations in their domain. Studying the observed behaviour results in a so-called protocol of actions. This protocol can be analysed to infer the actual problem-solving strategies used by the experts. Observation methods are most suitable for domains in which problem solving requires objects to be handled or overt actions to be performed.

The different methods in use for eliciting verbal reports are all variants of the interview. An interview may be focused but otherwise unstructured, with general questions; the topics to be addressed during the interview, but not the precise questions are prepared in advance. The interview may also be structured, containing mostly closed questions. An interview may be held orally or presented on paper, in the form of a questionnaire. Interviews are especially appropriate for domains in which it is relatively easy for experts to verbalise their knowledge, for example because their daily routines involve verbalisations of problem-solving behaviour. Unstructured, oral interviews closely resemble normal conversation. The experts are for example asked what they commonly do when they are confronted with a problem situation in their domain of expertise. This type of interview provides the interviewer with a global understanding of the structure of the knowledge domain, and of the type of strategy used for problem solving; more specifically, the interview results in an overview of the order in which the various reasoning steps generally are performed. Subsequent structured interviews are suitable to deepen the understanding, to further clarify the structure of the knowledge domain, and to zoom in on the details of the problem-solving strategies used by the experts.

Thinking aloud is the only method available for eliciting mental processes more or less directly. The experts are presented with a prototypical problem situation and are asked to verbalise their reasoning processes while solving the problem without interruption. Talking out loud concurrently with solving a problem situation leaves the experts no room or time for interpretation. They therefore directly reveal the strategies they use for solving the problem. Thinking aloud appears to be quite easy for most people and has been found not to interfere with their performance [14]. Think-aloud sessions are generally tape-recorded and transcribed for further analysis.

To conclude we would like to observe that all knowledge-acquisition methods require careful preparation by the knowledge engineer. While some prior knowledge of the domain is advisable on the one hand, to understand the experts' answers and to put the right questions, too much knowledge on the other hand may cause the engineer not to listen carefully to the expert and to

interpret answers in the light of his or her own views. Also, the interview questions and problem situations have to be prepared with care to guarantee that the knowledge aimed at is indeed elicited. Moreover, the setting in which the various sessions are to be held should be carefully selected. It should further be noted that, to the experts, some information or reasoning steps may be so self-evident that they are not verbalised, and consequently not acquired with any of the available methods. When analysing the acquired information, therefore, the knowledge engineer should be attentive to unjustified reasoning steps and prompt the experts for further elaboration.

The method

To acquire knowledge about the test-selection strategy used by our experts in the domain of oesophageal cancer, we employed an elicitation method that combined several of the methods reviewed above. Because the process of selecting diagnostic tests in essence is a type of problem solving in which the experts' behaviour is predominantly mental, the method of observation did not suit our purpose; with observation methods, we would not be able for example to gain insight in the experts' reasons for ordering specific tests. Also, thinking aloud with real, concurrent patients was not feasible, for evident reasons. We thus focused on interviews for eliciting verbal reports of problem-solving behaviour. We decided to conduct three consecutive interviews. The aim of the first, unstructured interview, was to obtain insight into the overall test-selection strategy employed and into the general arguments used by the experts. Since such an unstructured interview would be focused on the strategy and not on real patients, we were aware that we risked acquiring a general, text-book procedure rather than the experts' daily problem-solving routines. We decided therefore to follow up the first interview by a structured interview in which the experts were asked to think aloud while deciding, for a number of patients, which diagnostic tests to order. We felt that working with patient cases in a carefully conducted manner would closely fit in with the experts' problem-solving practice and would thus reduce possible biases from the elicitation method used. The aim of this second interview was to fill in details of the elicited test-selection strategy and, more specifically, of the arguments underlying the strategy. We decided not to work with historical patient cases, since the experts might recall these patients and let the real final outcomes influence their test-selection behaviour. We decided to employ fictitious patient cases instead, that were designed to be as realistic as possible. Working with fictitious patient cases brought the additional advantage that it allowed us to design cases with which we were able to explore the experts' decision boundaries. The third and last interview served to review and refine, if considered necessary, the flowchart that had resulted from the second interview.

The setting

The method for eliciting test-selection strategies outlined above was used with the two experts from the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, who had been involved in the construction of the decision-support system for oesophageal cancer from its very inception. Also present during the first two interviews were three researchers from the Institute of Information and Computing Sciences of Utrecht University; during the third and last interview, only two of the latter researchers participated. The first interviewer has a background in medical computer science, providing her with some knowledge about cancer in general and about the various diagnostic tests involved. The second interviewer has a background in mathematics. She had

constructed the decision-support system, for which she had held numerous interviews. The two domain experts and this interviewer thus were very well acquainted with one another. Over the years, moreover, the second interviewer had gained considerable knowledge about oesophageal cancer and its treatment. The third interviewer, to conclude, has a background in cognitive science and knowledge acquisition; she participated in the first two interviews only.

The three interviews were conducted at the Netherlands Cancer Institute, the home institute of the two experts. The first interviewer conducted the actual interview, asking the questions that had been prepared. We felt that the second interviewer, because of her accumulated knowledge about oesophageal cancer and its treatment, might unknowingly and unwillingly bias the experts in their answers. She therefore did not partake in the main interview and only asked the more elaborate questions about the experts' decision boundaries that emerged during the interviews. The third interviewer recorded the elicited knowledge and monitored the elicitation process. She typed the words from the interviews in a laptop, not just concentrating on relevant knowledge but also on remarkable meta-phrases uttered by the experts. We were aware that typing in a laptop was likely to result in a less accurate recording of the elicited verbalisations than taping with a voice recorder. Still a laptop was used instead of a voice recorder because the experts had previously indicated that they would feel embarrassed by the recording. They did not seem to feel uneasy by the use of the laptop.

6.3 The first interview

The first interview conducted with the two domain experts was an unstructured, oral interview. We briefly restate the main goal of the interview and the procedure followed, before presenting the results.

The goal
The goal of the first interview was to elicit general knowledge about the selection of diagnostic tests for patients suffering from oesophageal cancer. The main issues to be addressed during the interview were:

- Are the experts guided by a standard procedure for selecting diagnostic tests, or are they mainly guided by their own experience?
- Are the various tests performed in parallel or sequentially?
- Are some tests always performed together, or one after the other? For instance, is the biopsy always combined with a gastroscopic examination of the oesophagus?
- What are the criteria that the experts use for selecting tests?
- What are the experts' criteria to stop testing?

The procedure
For the interview, we prepared a small number of open questions. The main question was "Can you describe the way in which you select and order diagnostic tests, starting from the very first

consultation with a patient up to and including your final decision about the most suitable therapy?". Since we wanted to avoid biasing the experts, we let them talk freely and did not interrupt unless it was strictly necessary, for example when further elaboration was desired. These interruptions then only consisted of open questions such as "Why?" or "Can you describe what you are thinking right now?". The interview was conducted in the setting described in the previous section and took some 30 minutes.

The results
We found that upon first seeing a patient, the experts start with

- a physical examination of the patient;
- an interview with the patient, resulting in information about
  - the age of the patient;
  - the amount of weight loss suffered;
  - the patient's ability to swallow food.

Subsequently, independent of the results of the physical examination and interview, a number of diagnostic tests are ordered simultaneously:

- a gastroscopic examination of the oesophagus, resulting in information about
  - the shape of the primary tumour;
  - the location of the tumour in the oesophagus;
  - the circumference of the tumour;
  - the length of the tumour;
  - the presence of necrosis (substantial decay of tissue);
- a biopsy, mostly performed together with the gastroscopy, revealing
  - the histological type of the primary tumour.

In the sequel, we will refer to the physical examination, the interview, the gastroscopic examination and the biopsy together as the *starting package of tests*. The gastroscopic examination and biopsy serve to give insight in the presentation characteristics of the primary tumour. The physical examination and the interview with the patient result in an assessment of the patient's physical condition. We would like to note that, because our experts work at a highly specialised centre for cancer treatment, they generally see patients who are referred from regional hospitals where these tests have already been performed. Often, therefore, the test results are available. If the experts feel, however, that the tests were performed too long ago, they will order them to be performed anew.

After the results of the tests from the starting package have become available, the experts decide whether or not further testing is indicated. Patients with a very poor physical condition will now just receive highly palliative treatment, without having to undergo further testing. For all other patients, again a number of tests are ordered simultaneously:

- a CT-scan of the liver, lungs and thorax, resulting in information about
  - the presence of metastases in the loco-regional lymph nodes;
  - the presence of haematogenous metastases;
- an X-ray of the thorax, resulting in information about
  - the presence of haematogenous metastases in the lungs;
- a sonographic examination of the neck, providing information about
  - the presence of metastases in the lymph nodes in the neck;
- an endosonography of the local region of the primary tumour and of the mediastinum, giving insight in
  - the depth of invasion of the primary tumour into the oesophageal wall;
  - the presence of loco-regional lymphatic metastases.

These tests primarily serve to establish the extent of metastasis of the primary tumour. In the sequel we will refer to these four tests together as the *basic package of tests*. The tests from the basic package are again requested in parallel, but only after the results of the tests from the starting package have become available.

The remaining tests constitute the *extensive package of tests*:

- a bronchoscopy, resulting in information about
  - the depth of invasion of the primary tumour into the trachea and bronchi;
- a barium swallow with fluoroscopy, yielding insight in
  - the presence of a fistula (an open connection as a result of decay of tissue) between the oesophagus and the trachea;
- laparoscopic examinations of the liver, diaphragm and abdomen, resulting in information about
  - the depth of invasion of the primary tumour into the diaphragm;
  - the presence of haematogenous metastases in the liver;
  - the presence of metastases in the lymph nodes near the truncus coeliacus.

In contrast with the starting and basic packages of tests, not all tests from the extensive package are ordered just like that: one or more tests may be selected. Whether or not a specific test from the package is performed very much depends on the location of the primary tumour in the patient's oesophagus. If the tumour is located in the upper part of the oesophagus, a bronchoscopy and a barium swallow are performed to investigate whether or not the primary tumour has invaded the lungs. No laparoscopic procedures are performed, however, because the primary

tumour cannot have invaded the diaphragm and, moreover, it is very unlikely that lymphatic metastases will be found in the upper abdomen. If the primary tumour is located in the lower part of the oesophagus, on the other hand, no bronchoscopy or barium swallow are performed. Laparoscopic procedures are then ordered, yet only if surgical treatment is considered.

To summarise, we found that diagnostic tests are ordered in three different packages: the starting package, the basic package, and the extensive package of tests. The tests from the starting package will always be ordered, for every patient. These tests are performed in parallel. If a patient is in a very poor physical condition, the experts will then stop testing and provide highly palliative care. For all other patients, the basic package of tests is ordered as well. In addition, one or more tests may be chosen from the extensive package, dependent on the location of the tumour and on the most preferred therapy at that particular moment in the patient's management. Figure 6.1 presents a flowchart summarising the knowledge acquired from the first interview. In the flowchart, the boxes with rounded corners describe the beginning and end points of the experts' strategy. The rectangular boxes capture requests to perform specific tests; the diamonds represent alternative choices and capture the appropriate questions to decide upon which tests to order. The arrows in the chart indicate the sequential order in which the various decisions have to be taken and choices have to be made.

A discussion

From the first interview, various distinguishing features of the test-selection strategy employed by our experts emerged. We found that diagnostic tests are not ordered sequentially, where the decision whether or not to order a specific test depends on the result of a previous test. Instead, the tests are ordered simultaneously, in packages. The most important argument underlying the experts' strategy of parallel testing is the loss of time that would be incurred by sequential testing. It may take several weeks before the results of a test become available. The tumour may have progressed within that time and may thereby render the results from earlier tests obsolete. Moreover, patients often are in such a poor physical condition that it is preferable not to have them return to the hospital too often for yet another test. And, even more importantly, the loss of time may make the difference between a curable cancer and an incurable one. As a consequence of ordering diagnostic tests simultaneously, however, more tests are likely to be performed than are strictly necessary. When questioned about such unnecessary testing, our experts indicated they did not see it as a problem, as the tests are not inconvenient for patients and gaining time is of primary importance:

"Ordering tests in packages is time saving. You might perform too many tests, but time is so important that you order all the tests anyway."

Our experts' strategy of ordering diagnostic tests in packages thus is supported by a strong argument. This argument in fact indicates that a sequential test-selection strategy for our decision-support system would be an unacceptable oversimplification of problem-solving practice. Our system should offer a strategy that is able to select tests in packages, based upon the argument reviewed above.

A second feature that we noticed of our experts' test-selection strategy pertains to the role of the stage of a patient's cancer. The experts had indicated before that they first establish the most

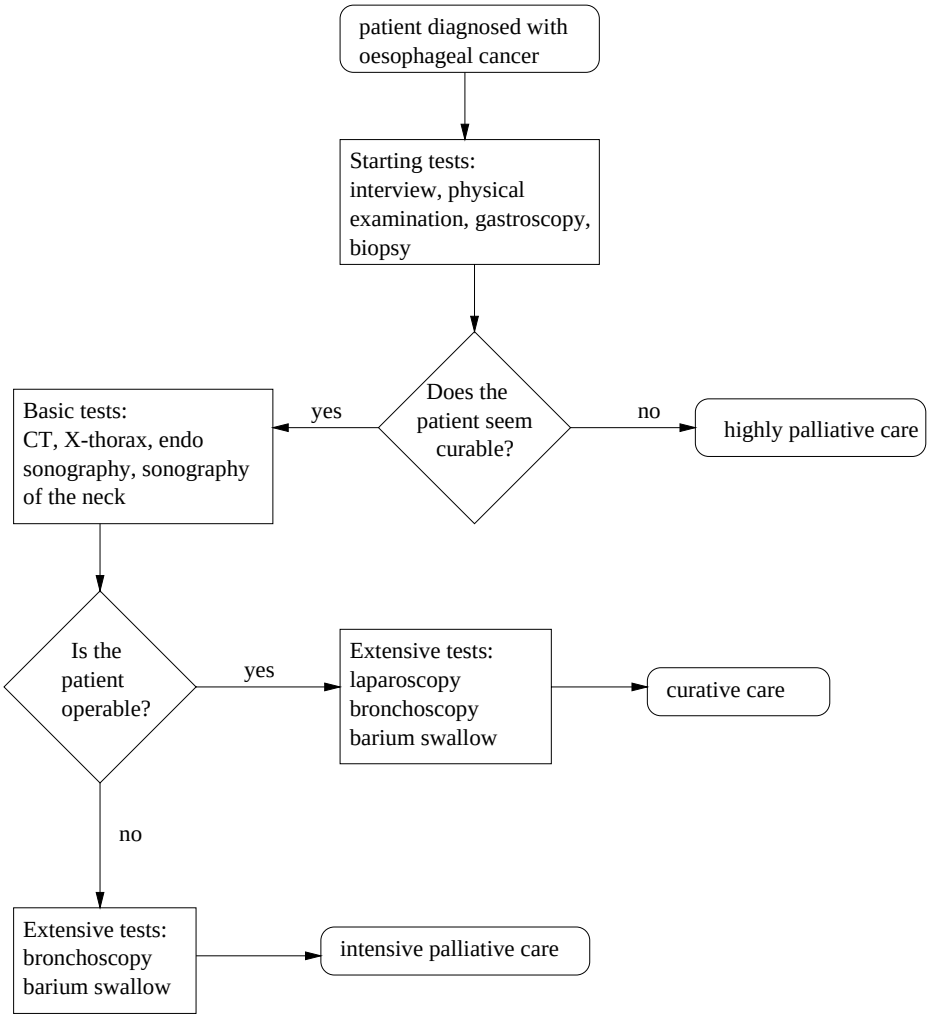


Figure 6.1: A flowchart summarising the knowledge from the first interview.

likely stage for a cancer before deciding upon an appropriate therapy. We found, however, that a cancer's stage only very indirectly plays a role in the selection of diagnostic tests. The decision which tests to order appears in fact to be based upon the experts' current idea about the most suitable therapy for a patient rather than on the uncertainty about the stage of his or her cancer. For example, if the tests from the basic package reveal lymphatic metastases in distant lymph nodes, then surgical removal of the primary tumour is no longer a feasible treatment option for the patient under consideration. Invasive laparoscopic procedures for establishing the exact extent of the cancer's metastasis then are not performed, even though the stage of the patient's cancer is still uncertain. To establish the most appropriate treatment alternative for a patient, the experts appear to gather information that helps them weigh the beneficial effects and complications to be expected for each alternative. Starting with the best possible alternative, that is to perform a surgical procedure, they order tests to see whether or not such a procedure is contra-indicated. Only if the test results indicate that a treatment alternative is not feasible, do they focus their attention on the next-best alternative. We note that over-treatment of a patient may easily prove to be fatal; under-treatment on the other hand may cause a patient to die prematurely from a cancer that might have been curable.

"You definitely do not want to miss anything. Missing and over-diagnosing are both bad. For physicians, however, missing feels worse than over-diagnosing as you want to do no harm. [...] When your feeling tells you something, but you cannot find it, you continue to look for it."

The test-selection strategy to be offered by our decision-support system should therefore take into account the impact of the possible results of a diagnostic test on the appropriateness of the available therapies rather than just its influence on the uncertainty of the most likely stage of a patient's cancer.

Closely related to the role of the stage of a patient's cancer is the role of the reliability characteristics of the various diagnostic tests in the test-selection strategy employed by our experts. As we have briefly mentioned in Chapter 4, the literature on medical decision making generally stresses the importance of taking the sensitivity and specificity characteristics of diagnostic tests into consideration upon test selection [58]. We found, however, that these characteristics play no role with our experts, since they do not gather information to reduce their uncertainty about the most likely stage of a patient's cancer. They also appear not to take the informativeness of the remaining available tests into account when they decide whether or not to stop testing. Their stopping criterion for ordering tests instead seems to be their certainty about the most appropriate treatment alternative for the patient at hand.

6.4 The second interview

From the first interview we had gained general insight in the test-selection strategy used by our two domain experts. Building upon the acquired knowledge, we followed up on the first interview with a second, more focused interview. Once again we briefly restate the goal of the

interview and the procedure followed, before presenting the main results.

The goal
The goal of the second interview was to fill in the details of the general test-selection strategy that had been acquired from the two experts during the first interview. More specifically, we aimed at eliciting the exact arguments used by the experts in their decisions to order a new package of tests or to refrain from further testing.

The procedure
For this second interview, we wanted to walk through the process of test selection for a specific patient, from the very first moment the experts see the patient up to and including the selection of the most suitable therapy. To this end, we designed a structured interview in which we carefully simulated the experts' daily problem-solving practice, by means of realistic patient cases. The experts were asked to think aloud while deciding for these patients which tests to order.

For the interview, we created eight fictitious patient cases; the specifics of these cases are summarised in Figure 6.2. Because we wanted to obtain as many details as possible about the test-selection strategy used by our experts, we created both patients for whom the most appropriate therapy seemed obvious and patients for whom the best therapy was not so evident. The first patient case mentioned in Figure 6.2 is an example of a patient for whom the most suitable therapy is quite evident. We expected that the experts would decide to administer highly palliative care for this patient. We further expected that the experts would not order the basic package of tests, nor any tests from the extensive package. For patient 6, however, it is not clear what the best treatment alternative would be. The patient has a moderately-sized primary tumour and is in a very good physical condition. In fact, from the results of the tests from the starting package, the patient appears to be curable. We therefore expected that the experts would consider surgery and would order the basic package of tests. The results of the tests from the basic package reveal distant metastases, unfortunately. With this evidence and the primary tumour being distal, we expected that the experts would refrain from further testing and would decide to administer intensive palliative care.

The eight patient cases were carefully designed as illustrated above, by means of the flowchart that had resulted from the first interview. For each patient case, moreover, we prepared a small number of questions that allowed us to more closely investigate the experts' exact decision boundaries. For patient 7, for example, we expected that the experts would pronounce him to be incurable, mainly as a consequence of his poor physical condition. The patient's tumour, however, is very small and seems to be resectable. To investigate under which conditions the patient would no longer be deemed incurable, we prepared various what-if questions, such as "What would you do if the patient were just 58 years of age?" and "What would you do if the patient were in moderate health?".

For each patient case we prepared three cards, or vignettes, with the results from the three different packages of tests. We created the three cards for every patient, even if it were very unlikely that even the basic package would be selected. The experts were informed that there were three cards per patient irrespective of whether or not we expected them to request the second package of tests or order tests from the extensive package. On each card, a result was indicated for each test from the package under consideration. Since the experts might feel that

Patient 1: An 87-year old male with a rather poor physical condition and a large primary tumour.

Patient 2: A 66 year old male who has a large primary tumour, yet is in good physical condition.

Patient 3: A 57-year old male having a very small primary tumour who is in an excellent physical condition. The tumour is located in the upper part of the oesophagus (a proximal tumour).

Patient 4: This 60-year old male is in excellent condition and has a very small primary tumour located in the lower part of the oesophagus (a distal tumour).

Patient 5: A 64-year old male in good physical condition with a moderately-sized primary tumour. The patient has lymph node metastases in his neck.

Patient 6: This 67-year old male has a moderately-sized distal primary tumour and is in a good physical condition. He has both proximal and distal lymphatic metastases.

Patient 7: An 80-year old male in a very poor physical condition, having a very small primary tumour.

Patient 8: A 59-year old male with a large primary tumour, in a very good physical condition.

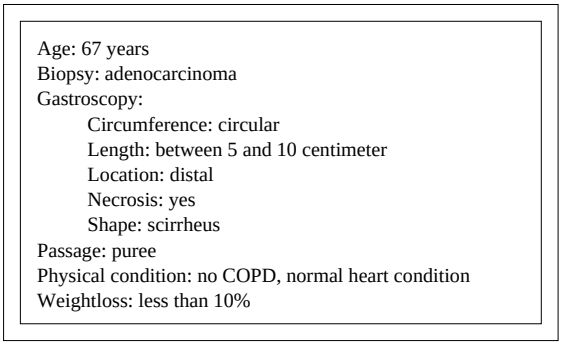
Figure 6.2: The eight patient cases designed for the second interview.

the first test results on a card are the most important, we decided to simply list the results in alphabetical order. We informed the experts about this alphabetical order to avoid biasing their problem-solving behaviour. Since the tests from a single package would be ordered in parallel, we decided to present the results of these tests simultaneously, and not one by one. As an example, Figure 6.3 shows the three cards that we prepared for patient 6.

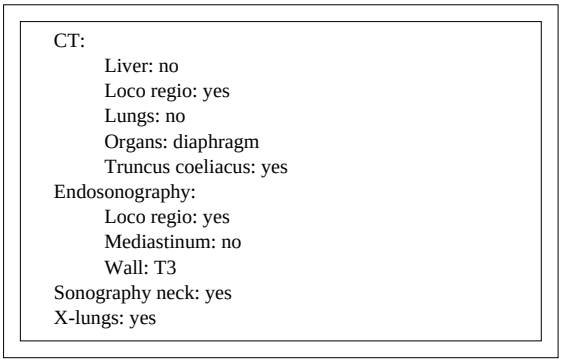
We asked the experts to discuss each patient case aloud. We asked them more specifically to verbalise their subsequent reasoning steps in ordering tests and to conclude the discussion of a patient case with an indication of the most appropriate therapy. We asked them to pretend that they were ordering real tests for real patients. For each patient, the card with the results of the tests from the starting package was presented first. Only when the experts indicated that they would order additional tests, would we show the second card with the test results from the basic package. Upon studying the second card, the experts also had access to the first card; they could thus survey the accumulated patient data. When still further testing was desired, the last card was presented. If the experts did not order any test from the extensive package or even from the basic package, the associated cards were not shown. The interview was conducted in the setting described in Section 6.3 and took approximately two hours for the eight patient cases.

The results

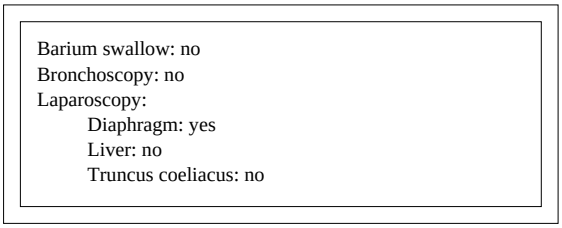
We present some fragments of the dialogue between the two experts while they were discussing patient 6, for whom the three cards shown in Figure 6.3 were created. Upon being presented with the first card, with the results of the tests from the starting package, the experts reasoned as follows:



(a)



(b)



(c)

Figure 6.3: The three cards representing the results of the tests from the starting package (a), the basic package of tests (b), and the extensive package (c), for patient 6.

"Oh, an average patient! We regularly see this type of patient. As his physical condition is quite good and the tumour is of moderate size, we might wish to consider surgery, so let's do the basic package of tests."

We presented the second card, with the results of the tests from the basic package:

"Mmm, there is some discrepancy here. Ah, well, we see them like this. Both proximal and distal metastases with a distal tumour. However, his condition is still quite good, so we would prefer to do something. We could consider palliative radiotherapy, also because he is not so old. [...] If the result from the sonography of the neck had been negative, there would only be metastases near the truncus coeliacus, which would make surgery a feasible option. Then, laparoscopy might be interesting. Now that the result of the sonography is positive, laparoscopy is no longer necessary."

The experts indicated that no further tests would be ordered.

The two experts discussed the eight patient cases at length. From the knowledge thus acquired, we constructed a new flowchart to capture the experts' test-selection strategy; the resulting flowchart is shown in Figure 6.4. To summarise, our observation from the first interview that diagnostic tests are ordered simultaneously in three different packages, was not contradicted by the second interview. Tests from the starting package are always performed. Only if a patient's physical condition is quite poor will the experts refrain from further testing and provide highly palliative care by positioning a prosthesis in the patient's oesophagus. For all other patients, the basic package of tests is ordered as well. If the results of the tests from this package reveal distant metastases, the experts will not order any new tests to be performed and start with local palliative care, that is, either a prosthesis is positioned in the patient's oesophagus or a palliative regime of radiotherapy is administered. If the metastases of the primary tumour are loco-regional, on the other hand, further tests will be selected from the extensive package to investigate whether or not surgical removal of the oesophagus is a feasible treatment alternative. If the tumour appears to be resectable, the appropriate tests from the extensive package will be performed. If it is evident that the tumour is not resectable, or if contra-indications for surgery have been found, such as a relatively poor heart condition, the laparoscopic procedures will not be ordered.

A discussion

From the first interview we had learned various distinguishing features of the test-selection strategy employed by our experts. The second interview served to corroborate and further detail our previous observations. It provided additional insight especially in the way in which the experts used the results from the different tests as arguments for their subsequent decisions. The second interview further most prominently demonstrated that, from the very first moment of seeing a patient, the experts think in terms of appropriate treatment alternatives. In fact, our observations strongly suggest that it is the task of selecting an optimal therapy that drives the ordering of diagnostic tests, rather than the task of establishing the stage of a patient's cancer. For example, when discussing one of the patient cases, the experts mentioned:

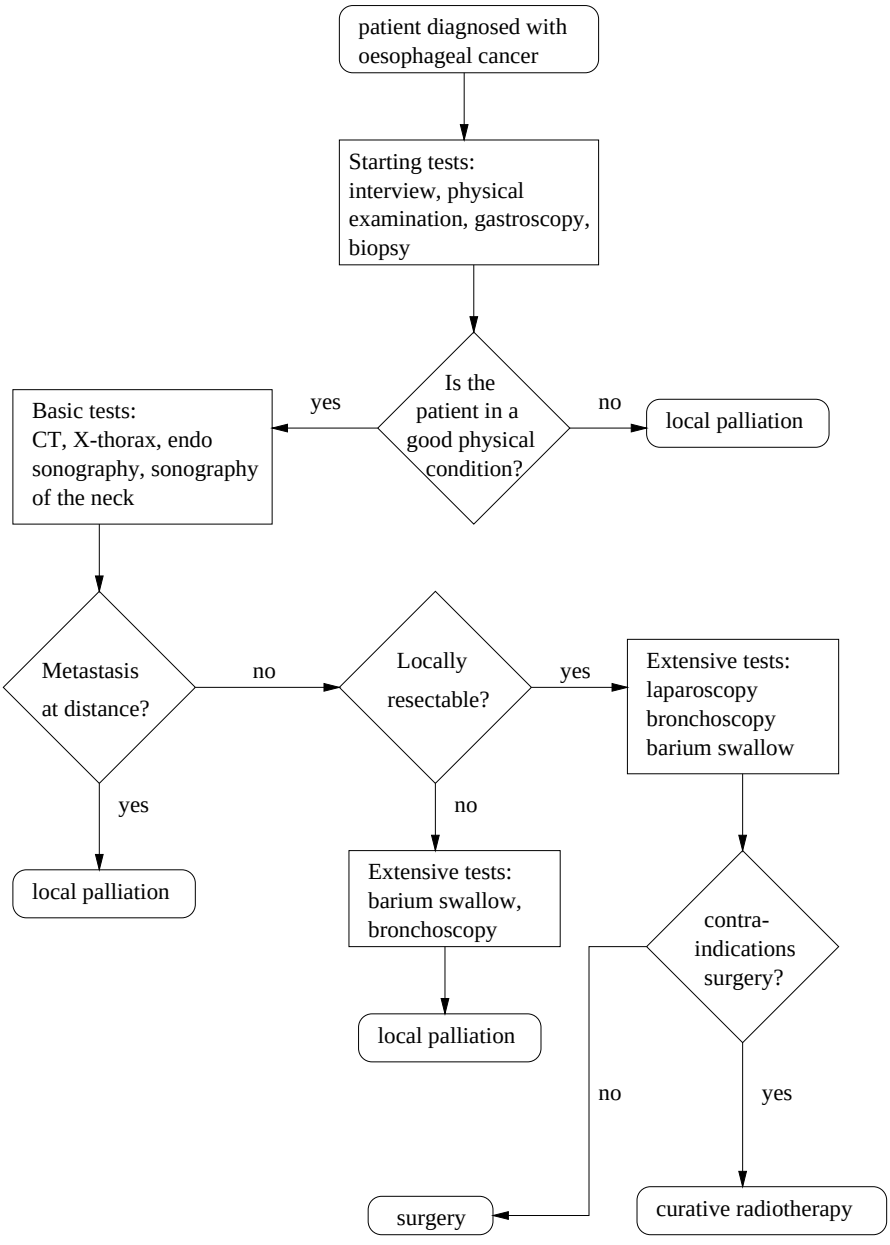


Figure 6.4: A flowchart summarising the knowledge from the second interview.

"We are thinking of surgery as the best therapy right now. Let's see how deep the tumour has invaded into neighbouring organs to see if surgery really is an option."

6.5 The third interview

From the second interview we had gained detailed insight in the test-selection strategy used by our two domain experts. Building upon the knowledge acquired from the first two interviews, we decided to follow up with a third interview to verify the exact decision moments captured in the flowchart that we had constructed. As before, we briefly restate the goal of the interview, before presenting the results that we obtained.

The goal
While the first two interviews focused mainly on acquiring the exact sequence of tests to be performed, the third interview focused on the precise decisions taken at the various different decisions moments in the process of test selection. Building upon the flowchart that had resulted from the second interview, the main issue to be addressed for each decision moment was whether it is indeed correct that that particular decision is taken at that particular moment in a patient's management.

The procedure
For the third interview, we presented to the experts the flowchart that we had constructed from the second interview. More specifically, we verbalised, in detail, the various different decision moments represented in the chart. We thus reviewed at length, with the two experts, each and every decision, beginning with the decision whether or not to order the starting package of tests and ending with the various different treatment alternatives. We asked the experts based on which test results the decisions were taken and what the implications of the decisions would be for the most suitable therapy. Prior to the interview, we had compared a number of patient data that had been made available to us for constructing and evaluating the oesophageal cancer network, against the flowchart. Upon doing so, we had noticed a number of apparent deviations in the data from the chart and decided to use those deviations to pose structured questions.

The results
As a result of the third and last interview, the flowchart that had resulted from the second interview was carefully refined. The modifications resulted mainly from the sharper decision criteria that the experts formulated when confronted with the chart. Also the change of focus from the sequence of diagnostic tests to the sequence of decisions and choice of therapy served to refine some of the modelled decisions. For example, in the original flowchart, the starting package of tests included the interview with the patient, a physical examination, a gastroscopic procedure, and a biopsy. From the data, we had noticed that when only the starting package of tests had been performed, for most patients an X-ray of the lungs and a barium swallow had been performed as well. When confronted with these observations, the two experts argued that an X-ray of the patient's lungs should indeed belong to the starting package of tests and not to the basic package, since this test will always be ordered for any patient regardless of his or her physical

condition. A similar conclusion did not hold for the barium swallow. The experts explained that this test had belonged to the starting package of tests for many years, but that, at present, it was no longer common practice to perform the test with every patient. The experts further indicated that it is common practice to perform laboratory examination of, for example, a patient's blood; this examination therefore was included in the starting package of tests.

The only fundamental modifications of the original flowchart pertain to the moment of establishing whether or not a patient's tumour is resectable. In the original chart, if no evidence of distant metastases had been found from the tests of the basic package, the next step was to decide whether or not the patient's tumour is locally resectable. While reviewing this decision moment the experts indicated that the result of a bronchoscopy can influence a physician's assessment of the resectability of the primary tumour. The bronchoscopy serves to give insight into whether or not the tumour has invaded the trachea; if it has invaded the trachea, the tumour has grown through the oesophageal wall into the trachea and is not resectable. In the flowchart, we moved the bronchoscopy from after the moment of deciding upon resectability to a moment prior to this decision. Note that the test should still only be performed on patients with a proximal tumour. The experts further indicated that a barium swallow is only performed if a patient's primary tumour has proved not to be locally resectable. A barium swallow serves to reveal the presence of a fistula. Only if the tumour has invaded the trachea, such a fistula may be present. The presence of a fistula forestalls the administration of any scheme of radiotherapy, and the only treatment option left is the positioning of an endoprosthesis. The details that we thus elicited, have been incorporated into the flowchart. The final flowchart is presented in Figure 6.5.

A discussion

In the first two interviews, we had learned a considerable amount of knowledge about the test-selection strategy employed by the experts in their daily routine. We had subsequently captured this knowledge into a flowchart. The focus in the first two interviews was on the sequence of diagnostic tests and on the packages in which the tests were ordered. By analysing the patient data that we had available, we found that we required more detailed information about the exact arguments, or in other words, the test results, based upon which the various decisions were taken. Since we would take these decisions as a basis for our automated test-selection facility, we decided to conduct a third interview to elicit the exact decisions and the underlying arguments. In this interview, we reviewed the flowchart that had resulted from the previous interviews. The majority of the changes resulting from the third interview were minor refinements of the chart with little impact on the modelled overall test-selection strategy. After the third interview, we felt confident that we had elicited the test-selection strategy employed by the two experts in sufficient detail. We continue our discussion based upon the flowchart presented in Figure 6.5.

6.6 Comparing data against the elicited strategy

Since we would use the arguments underlying our experts' problem-solving practice to build a facility for automated test selection, we wished to ascertain that we had elicited their test-selection strategy at a sufficient level of detail. To this end, we studied the medical records of real patients from the experts' institute and compared these against the flowchart that had resulted

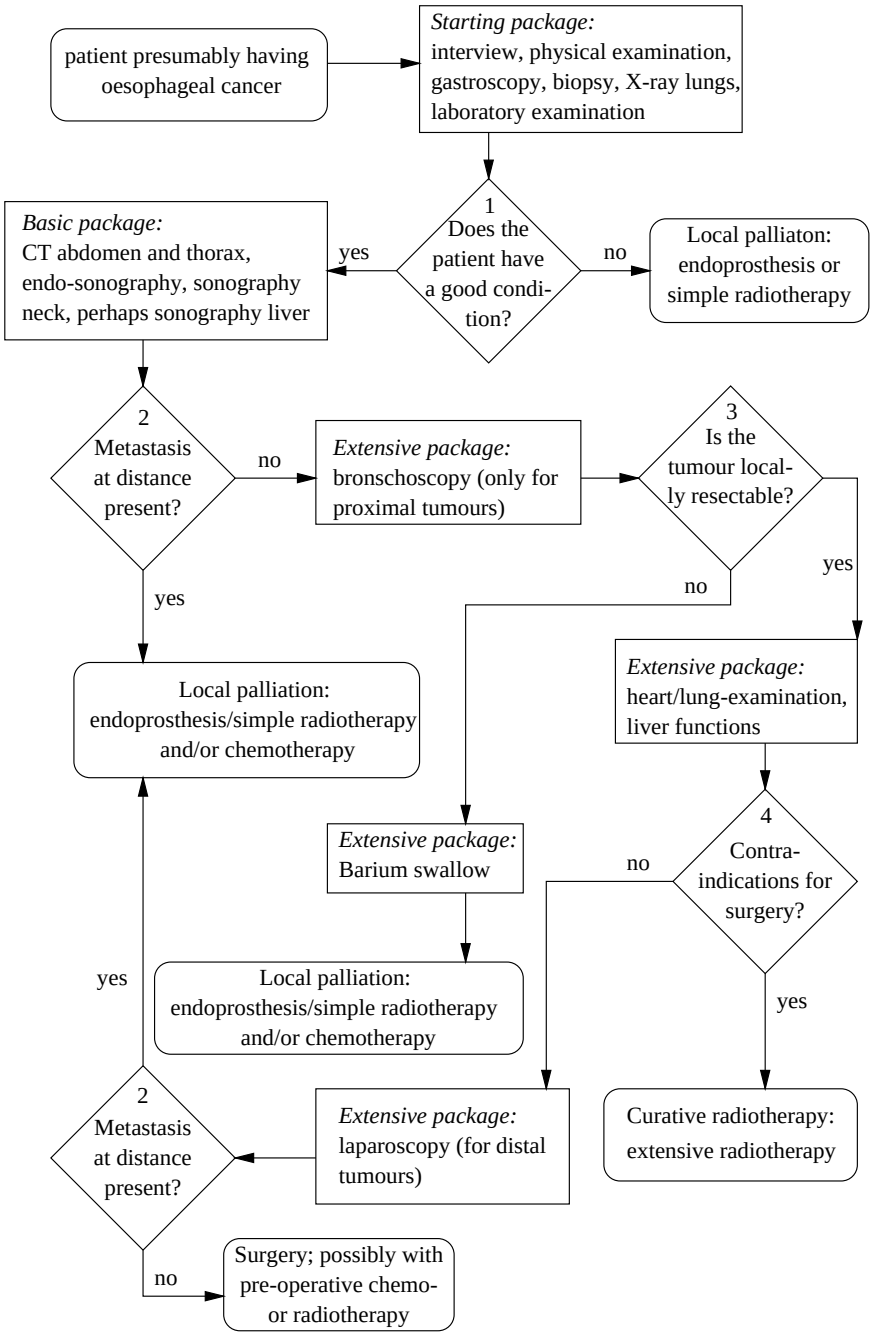


Figure 6.5: The flowchart resulting after the third interview.

after the third interview. We found that the results of the comparison were influenced by various issues. For example, changes in medical practice over time and patient data originating from referring hospitals are likely to have obscured the results to a considerable extent. Missing values, moreover, seriously hampered the comparison. Since many retrospective databases in medicine are likely to have the same limitations as the data under study, we feel that the problems that we encountered upon our analysis of the data, are likely to be found in other medical applications as well.

6.6.1 The data

For comparing the elicited test-selection strategy against our expert's problem-solving practice, we had at our disposal the medical records of 156 patients diagnosed with oesophageal cancer from the Antoni van Leeuwenhoekhuis. The data were collected over a period of 11 years. The data had been entered into a database from paper records, by two medical students. For each patient, various symptoms and test results were available, such as the results from a gastroscopic examination of the oesophagus and an assessment of the patient's ability to swallow food. For each patient, a maximum number of 28 results could have been available, originating from 12 different physical tests. The number of available results per patient, however, ranges between 8 and 24. On average, 13.9 tests results are available per patient; the median of the number of available results equals 14. The data therefore are relatively sparse. We would like to note that the data only record the results from the various tests and do not provide any information about *when* the tests were performed. For each patient, also various other data were recorded. For example, the stage of the patient's cancer, as established by the attending physician, as well as the administered therapy are available. In addition, values for various intermediate, unobservable variables are stated such as the absence or presence of haematogenous metastases; these values basically are conjectures by the physician.

6.6.2 Analysis of the patient data

We studied the collection of available patient data and compared these against our flowchart. For every patient, we established from the flowchart which decisions should have been taken and which tests should have been performed, and compared these against the data recorded for the patient. We will discuss our results in three parts. In the first part, we focus on the availability of results of the various different diagnostic tests and packages. The second part addresses whether or not packages were ordered in accordance with the flowchart, based upon the actually available test results. Due to the large number of missing values, we were not able to establish the exact decisions for many patients. Since we do not only have available actual test results but also the physicians' conjectures about the values of various intermediate, unobservable variables, we decided to use these conjectures in a second evaluation of whether or not packages were ordered in accordance with our flowchart. The results of this second evaluation are reported in the third part.

Availability of test results

We begin by studying the availability of results from the various different tests. From the

flowchart, we have that every patient has to undergo all tests from the starting package. However, as shown in Table 6.1, the ability to swallow food, the weightloss suffered and the WHO-status were not recorded for every patient. We further observe that the gastroscopic examination indeed was performed for all 156 patients. However, for some patients, the results from this examination with respect to the shape and circumference of the primary tumour and the presence or absence of necrosis were not recorded. These missing values are most probably due to the presence of a large tumour: if the endoscope cannot pass by the primary tumour, the laboratory technician cannot assess the tumour's characteristics. From the 13 possible test results from the starting package of tests, between 7 and 11 were available per patient. Of the five physical diagnostic tests, between two and four were performed. A test was marked as having been performed on a patient if at least one of its results was recorded in the patient's data. We would like to note that from the starting package of tests, the results of the physical examination of the patient's neck, for example, were not recorded in our data at all.

From the flowchart, we have that all patients with a good physical condition should undergo the basic package of tests; the patients with a poor condition should not receive any further testing. For 21 (or 13%) of the 156 patients, we could not establish the physical condition from the available data. For 19 of these patients, the WHO-status was not recorded; for the other two patients, the amount of weightloss suffered was missing. 125 patients (80%) were in a good physical condition. 10 patients (6%) had a poor condition. We observe that one (10%) of these 10 patients indeed did not undergo further testing; 9 of these 10 patients (90%) received the basic package of tests although they should not have received any further tests according to the chart. From the 21 patients for whom we were not able to decide upon their physical condition, 19 patients (90%) received tests from the basic package; the remaining two patients (10%) did not receive any further tests. Of the 125 patients with a good physical condition, 13 patients (10%) did not undergo any test from the basic package of tests. The remaining 112 patients (90%) indeed received the package. We marked a package as having been performed if one or more results for any of the tests were recorded in the patient's data. We found that 24 patients of the total of 140 patients who received the basic package, underwent all four diagnostic tests from the package. Of the 16 patients who did not receive the basic package of tests, the medical records of two patients revealed that the X-ray of the lungs showed metastases. It therefore is readily explained why the basic package of tests had not been ordered for these patients.

From the flowchart, we now observe that a bronchoscopy should only be performed on patients with a good physical condition who have a proximal primary tumour that has not yet resulted in distant metastases. For one patient (0.6%), all conditions for a bronchoscopy were fulfilled. For 148 patients (95%), at least one of the conditions were violated. For the remaining 7 patients (4%), we could not decide whether or not they should undergo the bronchoscopy. The one patient who fulfilled all conditions did not receive the bronchoscopy. For the 148 patients who should not undergo a bronchoscopy, 13 patients (8%) nevertheless received one. Eleven of these patients had a non-proximal tumour. Of the 7 patients for whom we could not establish whether or not they should undergo a bronchoscopy, 4 patients received one.

From the flowchart, we further have that a barium swallow should only be ordered for patients in a good physical condition who have a non-resectable primary tumour without any distant metastases. For three patients (2%), all conditions for a barium swallow were fulfilled. For 69

patients (44%), at least one of the conditions were violated. For the remaining 84 patients (54%), we could not decide whether or not they should undergo the barium swallow. The three patients who fulfilled all conditions, indeed underwent a barium swallow (100%). From the 69 patients for whom a barium swallow should not be ordered, 54 patients (78%) nevertheless received one. From the 84 patients for whom we did not have sufficient data available to decide upon the barium swallow, 71 patients (85%) underwent the test.

To conclude, we address the availability of results from a laparoscopic examination. From the flowchart, we have that a laparoscopic examination should only be ordered for patients in a good physical condition who have a resectable primary tumour without any distant metastases and for whom no contra-indications for surgery are recorded. For 48 patients (31%), at least one of the conditions were violated. For the remaining 108 patients (69%), we could not decide whether or not they should undergo the laparoscopic examination, because of lack of information about the possible contra-indications for surgery. Of the 48 patients who should not receive a laparoscopy according to our flowchart, 12 patients (25%) nevertheless underwent one. Of the 108 patients for whom we could not decide whether or not they should receive a laparoscopy, 14 patients underwent one. We observed that 14 patients received both a barium swallow and a laparoscopy, although our flowchart indicates that these two tests should never be ordered for the same patient.

Analysis of the decisions

After having studied the overall availability of test results, we now conduct a more detailed walk-through of the flowchart, focusing on the various decisions. For each decision moment in the chart, we analyse, more specifically, which decision had to be taken. A summary of the results is given in Figure 6.6 for ease of reference.

From the flowchart we observe that every patient should undergo the starting package of diagnostic tests. After the results from the tests of this package have become available, the physician has to decide upon the physical condition of the patient. The condition is considered to be good if the patient has a WHO status smaller than 2; if the WHO status is larger than 2, the patient's condition is considered poor. A patient with a WHO status of 2 is in a poor condition if he is over 85 years of age and has suffered more than 15% weightloss; otherwise, his condition is taken to be good. If his physical condition is poor, the patient should undergo local palliation without any further testing; otherwise, the patient will receive the basic package of diagnostic tests. As discussed above, we found that, of the total of 156 patients, 125 patients (80%) were in a good physical condition. Ten patients (6%) were in a poor physical condition. For 21 patients (13%) we were not able to decide upon their physical condition. We found that, of the 10 patients who should not undergo further testing, only one patient indeed did not receive any further tests. The other 9 patients received further tests from the basic package of tests. These nine patients in addition received a barium swallow from the extensive package of tests. One of these 9 patients even underwent a bronchoscopy and three of these patients received a laparoscopy.

For the 125 patients with a good physical condition and the 21 patients for whom we could not decide upon their condition, we continue the comparison against the flowchart. From the chart we now observe that, once the results of the tests from the basic package are available, it is again decided whether the patient should undergo further testing or receive local palliation. This second decision is based upon the absence or presence of distant metastases. If metastases distant from the primary tumour have been found with at least one of the tests, the patient should

Table 6.1: Availability of test results.

Test	% *available (n=156)*
Passage	98.7
Weight loss	95.5
Age	100
WHO-status	87.8
Gastro-location	100
Gastro-shape	99.4
Gastro-length	100
Gastro-circumference	98.7
Gastro-necrosis	98.72
Barium swallow	82.1
Biopsy	99.4
Bronchoscopy	10.9
X-ray lungs	75
Sonography neck	21.2
Sonography liver	55.8
Endosonography wall	30.8
Endosonography mediastinum	30.8
Endosonography loco	30.8
CT-organs	74.4
CT-lungs	72.4
CT-local region	71.2
CT-liver	62.2
Lapa-diaphragm	10.9
Lapa-liver	12.2
Lapa-truncus coeliacus	11.5

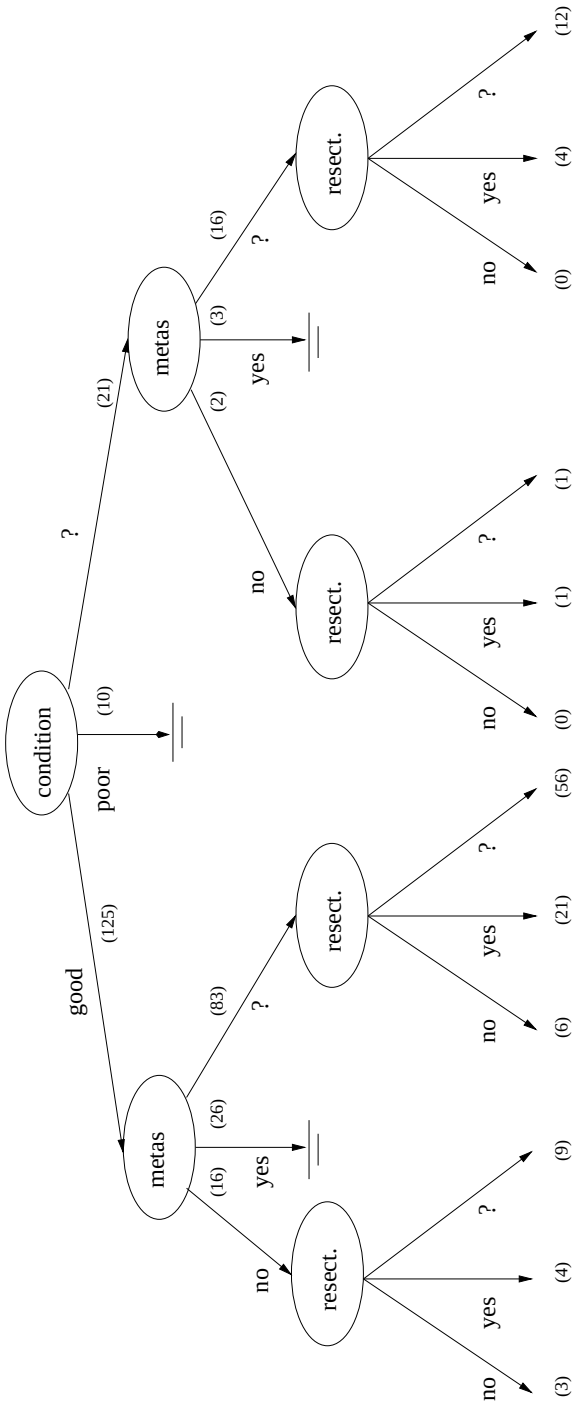


Figure 6.6: A tree summarising for how many patients which decision could be made according to our flowchart, based upon just the available results.

receive local palliation without further testing; if there is no evidence of distant metastases, additional tests should be performed from the extensive package. For proximal tumours, all metastases except those located in the cervical lymph nodes, are classified as being distant; for distal tumours, all metastases except those in the coeliac lymph nodes are distant; for a tumour located in the mid part of the oesophagus, all metastases are distant. We found that, of the 125 patients with a good physical condition, 16 patients (13%) did not have any evidence of distant metastases; these patients therefore should undergo further testing. Distant metastases were established in 26 patients (21%). Although these patients should receive palliative treatment without further testing, 20 patients of these 26 received one or more tests from the extensive package. For 83 patients (66%), unfortunately, we could not establish the presence or absence of distant metastases from the available test results. Of the 21 patients for whom we could not decide upon their physical condition, two patients (10%) had no evidence of metastases distant from the primary tumour. In three patients (14%), distant metastases were established; these patients nevertheless received the barium swallow from the extensive package of tests. For the remaining 16 patients (76%), we did not have sufficient data to decide upon the absence or presence of distant metastases.

We continue our walkthrough of the flowchart with the 117 patients for whom we were able to decide that no metastases distant from the primary tumour were present, or for whom we could not decide upon the absence or presence of distant metastases. In the absence of distant metastases, only patients with a proximal tumour should receive a bronchoscopy from the extensive package of tests. Of the 18 patients for whom further testing was indicated, only one patient (6%) had a proximal tumour. This patient did not receive a bronchoscopy, however. From among the other 17 patients, one patient received a bronchoscopy even though his primary tumour was located in the mid part of the oesophagus. When questioned about this observation during the third interview, the experts indicated that large tumours in the middle of the oesophagus can also grow into the trachea. The bronchoscopy is then performed to establish whether or not the trachea has actually been invaded. We concluded that our flowchart did not capture the criterion to decide upon performing a bronchoscopy in sufficient detail and consequently adapted it. Of the total of 99 patients for whom we were not able to decide upon the absence or presence of distant metastases, 13 patients (13%) received a bronchoscopy. Nine patients of these 13 did not have a proximal tumour. We would like to note that 13 of all 156 patients received a bronchoscopy while our flowchart indicated that they should not have; 11 of these 13 patients had a tumour in the middle or lower part of the oesophagus.

The flowchart now shows that for the remaining 117 patients under study, it is decided whether or not the primary tumour is locally resectable. This decision is based upon the depth of invasion of the tumour. If the primary tumour has grown through all the layers of the oesophageal wall and into a neighbouring organ, it cannot be locally removed; if the tumour is still confined within the oesophageal wall, it is considered to be resectable. The depth of invasion of the primary tumour is determined from the result of the endosonography of the oesophageal wall. If the result is T4, there is evidence that the tumour has invaded a neighbouring organ and, hence, is not resectable; otherwise, the tumour is restricted to one of the inner layers of the oesophageal wall and is resectable. Upon studying the medical records of the 16 patients in a good physical condition who did not have any evidence of distant metastases, we found that 4

patients (25%) had a resectable tumour. The tumours of three of these 16 patients patients (19%) were not resectable. For the remaining nine patients (56%), resectability of their primary tumour could not be established from the available data because no result from an endosonography of the oesophageal wall had been recorded. For the 83 patients who were in a good physical condition, but for whom we were not able to decide upon the absence or presence of distant metastases, 21 patients (25%) had a resectable tumour. The tumours of six patients (7%) were not resectable. For 56 of these 83 patients (67%), resectability of their tumour could not be established. Of the two patients for whom we could not decide upon their physical condition yet who did not have any distant metastases, one patient (50%) had a resectable tumour; for the other patient, we could not decide upon the resectability. For the 16 patients for whom we could not decide upon their physical condition nor upon the absence or presence of distant metastases, four patients (25%) had a resectable tumour; for the other twelve patients (75%), we could not decide upon the resectability of their primary tumour.

We continue our analysis with the 87 patients who do not have a locally resectable tumour, no distant metastases, and are in a good physical condition, or in whom one or more of these conditions could not be established from the available data. For just three of these 87 patients, all conditions for a barium swallow were fulfilled; all three patients indeed underwent the test. For 78 of the remaining 84 patients (93%), were were not able to decide upon the resectability of the tumour; 65 of these 78 patients (83%) underwent a barium swallow. The other 6 of these 84 patients (7%) had an irresectable tumour, however; 5 of these 6 patients (83%) received a barium swallow. It is remarkable that, of the total of 156 patients, 54 patients received a barium swallow while their tumours appeared to be resectable or had given rise to distant metastases, or they were in a poor physical condition. These 54 patients therefore underwent the test while it was not indicated by the flowchart. When questioned about this observation, the experts indicated that in the past, the barium swallow was used as the very first test upon a suspicion of cancer of the oesophagus. Nowadays, the barium swallow is no longer part of the starting package of diagnostic tests. The experts further indicated that a barium swallow is likely to be performed when the patient is considered for radiotherapy. The test then is not performed as part of the diagnostic phase of the management of the patient, but for therapeutic reasons. These two considerations are likely to explain the difference we observed between the flowchart and the patient records.

To conclude, the flowchart indicates that for each patient in a good physical condition with a resectable primary tumour that has not yet resulted in distant metastases, it is decided whether or not surgery is feasible. This decision is based upon an assessment of the condition of the patient's heart, lungs and liver. Since we did not have such assessments available in our data collection, unfortunately, we were not able to complete the comparison of our data against the flowchart.

The analysis reviewed

From the previous section, we observe that our comparison of the available patient data against the flowchart was seriously hampered by the large number of missing test results: for 113 of the total of 156 patients (72%) we were not able to establish the correct decision based upon the results of the various different diagnostics tests, at one or more decision moments in the flowchart (disregarding the decision pertaining to the contra-indications for surgery). Now, as we have described in Section 6.6.1, our data collection contains not just results from diagnostic tests, but also conjectures by the attending physicians. These conjectures pertain for instance to

the true presence or absence of metastases in various lymph nodes and organs. We decided to conduct another walkthrough of the flowchart, this time using both the available test results and the recorded conjectures. The results are summarised in Figure 6.7, for ease of reference.

From the flowchart we once again have that the first decision concerns the physical condition of a patient. We recall that for 21 patients we could not decide upon their condition. Since none of the recorded conjectures provides further information about a patient's condition, we still could not decide upon the physical condition of these 21 patients. For the 125 patients who proved to have a good physical condition, we previously found that for 83 of these patients we could not decide upon the presence or absence of distant metastases. Especially the results from a physical examination of the patient's neck and from a CT-scan of the truncus coeliacus were often missing for this purpose. Using the physicians' conjectures with regard to the absence or presence of metastases near the truncus coeliacus and in the neck, we were now able to decide whether or not distant metastases were present for 36 of these 83 patients. For 12 of these 36 patients (33%), no metastases distant from the primary tumour were indicated; for the other 24 patients (67%), such metastases appeared to be present. In our previous walkthrough, we further found that for 16 patients we were not able to decide upon their physical condition nor upon the absence or presence of distant metastases. For 5 of these patients, we still were not able to decide upon the absence or presence of distant metastases with the help of the recorded conjectures. For 4 of the remaining 11 patients (36%), the conjectures indicated that no distant metastases were present. For 7 of the 11 patients (64%), distant metastases were present.

We now continue our walkthrough with the 86 patients who are in a good physical condition and for whom we were able to decide that no distant metastases were present, or for whom we could not decide upon either the condition or the absence or presence of distant metastases. We consider for these patients the decision with respect to the resectability of their tumour. We recall that this decision is based upon the result of an endosonography of the oesophageal wall. Unfortunately, as we mentioned before, this result was missing rather often. The T stage of the tumour, moreover, was not summarised in any conjecture by the physicians. Using the conjectures therefore did not further contribute to our previous analysis.

Using the conjectures recorded in the patient data, we reduced the number of patients for whom we were not able to take one or more decisions from our flowchart, from 113 (72%) to 83 (53%). Note that for still quite a large number of patients, we could not complete our walkthrough of the flowchart as a consequence of missing values.

6.7 Conclusions

Upon working with our decision-support system for oesophageal cancer, we felt that using the sequential test-selection strategies commonly proposed in the decision-making literature would be an oversimplification of our experts' daily problem-solving practice. We decided to acquire knowledge about the actual strategy used by the experts to provide for the design of a more tailored test-selection facility. For this purpose we used an elicitation method that was composed of three focused interviews: an unstructured interview, followed up by a structured one and an interview for refinement purposes. With the first, unstructured interview, we found that tests were ordered not sequentially but in three different packages, with the tests from a single pack-

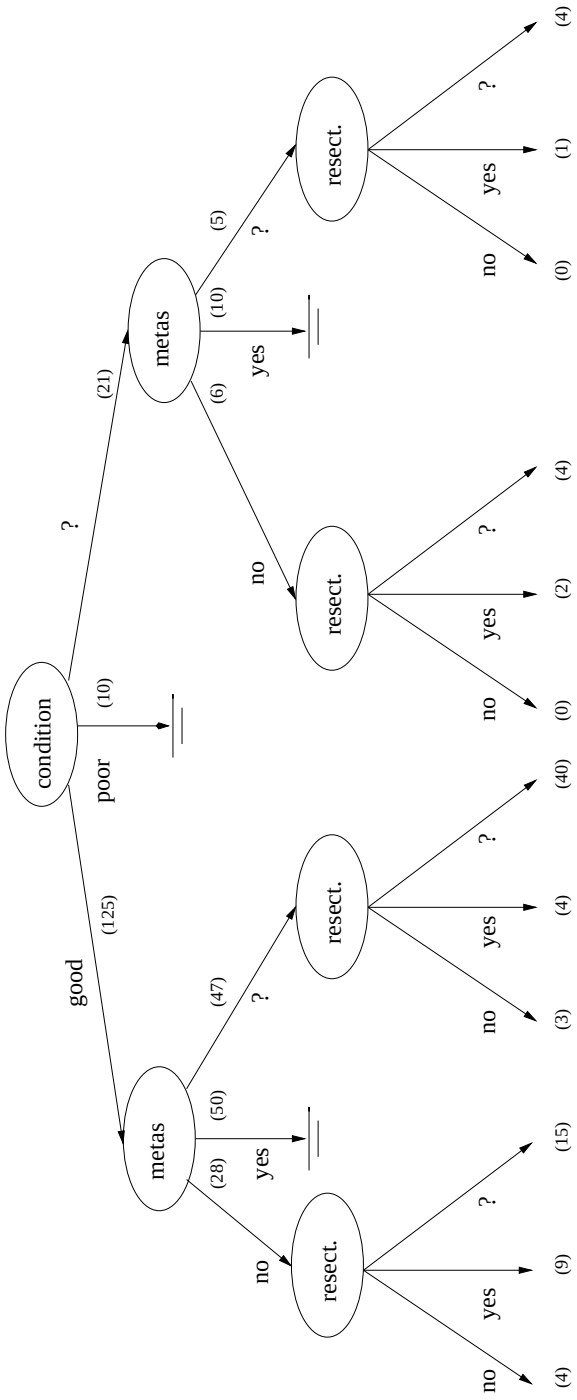


Figure 6.7: A tree summarising for how many patients which decision could be made according to our flowchart, based upon the available test results and recorded conjectures.

age ordered simultaneously. With the second, highly structured interview, we were able to fill in the details of the general strategy that we had elicited. More specifically, we were able to establish the different arguments underlying the experts' test-selection decisions. The third interview aimed at reviewing the flowchart resulting from the second interview in detail and refining it.

With the structured interview, we carefully simulated the experts' problem-solving practice through the use of fictitious patient cases. Each patient case was captured by three different cards, or vignettes, with the results of the three packages of tests. These cards were presented sequentially to the experts. We found that our approach indeed closely fitted in with our experts' daily practice. In fact, they explicitly mentioned that using the cards was very intuitive:

"These cards and the way we discuss them are very similar to how patients are presented during the sessions we have with colleagues when we discuss patients."

The use of the cards thus worked quite well. The only difference with the experts' daily problem-solving practice may have been that in the current interview setting, not facing a real patient and with less time pressure, the experts were more consistent and more thorough in their decisions than they usually are. One of the experts mentioned:

"Perhaps we are now more consistent than we normally are in practice."

We felt that linking up with practice was highly advantageous for the purpose of acquiring knowledge of the test-selection strategy used by our experts. We would like to note, however, that especially for the set-up of the second, structured interview, prior domain knowledge appeared to be imperative. Without considerable prior knowledge, we would not have been able to design the fictitious patient cases in a way that allowed us to explore the experts' decision boundaries.

We feel that a test-selection facility offered by a decision-support system should support physicians in their daily problem-solving practice and should therefore be based upon the arguments experts use in their decisions to order specific tests. We feel, in addition, that to design such a facility, knowledge about the actual test-selection strategy used should be elicited from experts in the domain of application. A standard sequential facility may then turn out to be unacceptable to the physicians who are the projected users of the system. In this chapter, we have demonstrated that eliciting test-selection knowledge from experts can indeed be feasible and is likely to result in a wealth of detailed information that can provide for a carefully tailored test-selection facility.

Since we wished to ascertain that we had elicited sufficient detail for the design of an automated test-selection facility, we compared the medical records of 156 patients against the elicited strategy, using the flowchart that we had constructed from the interviews. We found that the results of our comparison were influenced to a large extent by various issues. The data in a patient's record often originated from different hospitals. The data being available from different sources tended to influence the experts' selection of tests and thereby obscured their compliance with the flowchart. Moreover, our data was collected over a longer period of time during which the medical practice had changed. These changes were hard to detect from the data itself without consulting the experts. To conclude, our analysis was seriously hampered by a relatively large

number of missing values, especially since a missing value could have different meanings. As a consequence of these issues, we were not able to derive any definite conclusions from our comparison. We feel that the above and other related issues are not reserved for our field of application. In fact, we feel that these issues are likely to arise in many retrospective real-life medical data collections and should be carefully analysed.

CHAPTER 7

New algorithms for test selection in probabilistic networks

In Chapter 3, we reviewed the various concepts that are involved in the design of an automated test-selection facility for a decision-support system in general. We argued that, roughly speaking, such a facility is composed of three basic components: an information measure for capturing the uncertainty about the overall diagnosis, an actual test-selection loop, and an associated criterion for deciding when to stop gathering further information. In the context of a probabilistic network, the information measure is typically defined on the probability distribution over the main diagnostic variable of the network under study. For the oesophageal cancer network, for example, the information measure is defined on the probability distribution over the variable representing the stage of the oesophageal cancer. In Chapter 5 we analysed the most often employed information measures from a fundamental perspective as well as from an experimental perspective. We concluded that, for test-selection purposes, both the Shannon entropy and the Gini index are suitable information measures for use in a test-selection facility.

With respect to the test-selection loop, we argued in Chapter 3 that most algorithms in use in practical decision-support systems serve to select diagnostic tests myopically. In each iteration of the loop, the most informative test variable is selected from among all possible test variables to indicate the next test to perform. The user is then prompted for the value of the selected variable. The result is entered into the probabilistic network of the system and propagated to establish the posterior probabilities for all variables. From the set of test variables still available, the next variable is selected. This process of selecting test variables and propagating their results is continued until a stopping criterion is met or until results for all test variables have been entered into the network.

We feel that the test-selection strategy that is induced by a myopic algorithm is an oversimplification of the experts' problem-solving strategy in many medical fields of application. From the

three interviews that we held with our experts in the domain of oesophageal cancer, as reported in Chapter 6, for example, we learned, that the test-selection strategy they employ in their daily practice is aimed at deciding between different therapies rather than at establishing the stage of the disease. In their strategy, moreover, different subgoals can be identified that are addressed sequentially; these subgoals are captured by the decision moments in our flowchart from Figure 6.5. We feel that a more involved test-selection facility should be able to take such subgoals into account.

In addition to the considerations above, we feel that also the myopic nature of the algorithms currently in use presents an oversimplification of test selection in many fields in medicine. In the domain of oesophageal cancer, for example, multiple test variables serve to model the results of a single physical test. Moreover, our experts have been found to order tests in packages to reduce the length in time of the diagnostic phase of a patient's management. For the latter purpose, especially, a non-myopic test-selection algorithm would be required. We argued in Chapter 3, however, that a fully non-myopic algorithm is computationally very demanding and, in fact, may easily prove to be too much demanding for practical purposes. With respect to the first consideration for non-myopia, we observe that an algorithm that takes a fixed clustering of test variables into account, would suffice. Such an algorithm would retain some of the idea of non-myopia, yet stay computationally feasible.

In Chapter 3, we argued that a test-selection facility should result in a strategy in which not too many tests are selected. Even more importantly, it should prevent the test-selection process from halting too soon. To prevent overtesting, a test-selection algorithm is typically extended with a stopping criterion. With such a criterion the algorithm computes if performing more diagnostic tests is necessary or that the physician can safely stop testing. In the literature on automated test selection, relatively little attention has been paid to stopping criteria. As a stopping criterion, generally the probability of the most likely value of the main diagnostic variable is examined: if this probability exceeds a predefined threshold, then the selection of further tests is halted.

In this chapter, we address the issues of employing subgoals and of including restricted concepts of non-myopia in test-selection algorithms. Section 7.1 presents three new algorithms that build upon these issues. In Section 7.2 we introduce four stopping criteria to be used with our new algorithms. In Section 7.3 we describe the experiments that we conducted with our new algorithms and stopping criteria. This chapter ends with our conclusions in Section 7.4.

7.1 New algorithms for test selection

Before presenting our new algorithms for test selection in probabilistic networks, we briefly review the basic myopic algorithm discussed in Chapter 3. We recall that this algorithm takes for its input a set $\mathcal{T}$ of test variables. For its output, the algorithm sequentially prompts the user to supply a value for a selected variable $T_i \in \mathcal{T}$. The value entered by the user then is propagated through the probabilistic network under consideration before the next variable is selected and presented to the user. The selection of a next test variable is based upon the largest expected decrease in the value of the information measure used. Based upon our considerations from Chapter 5, we assume from now on that the algorithm employs the Gini index of the probability

distribution over the disease variable; note that any information measure, however, can be used. The basic myopic algorithm, including a stopping criterion, now amounts to the following in pseudo-code:

Myopic test selection

input:
$\mathcal{T}$: list of test variables T_i

Stop = *false*
while $\mathcal{T} \neq \varnothing$ and *Stop* ≠ *true* **do**
    compute most informative $T_i \in \mathcal{T}$
    prompt for evidence for T_i and propagate
    remove T_i from $\mathcal{T}$
    compute *Stop*
od

From a computational point of view, the most expensive step in the algorithm is the step in which the most informative test variable is selected. In this step, the Gini index $G(\Pr(D \mid T_i))$ has to be computed for each test variable T_i. We note that the probability distributions $\Pr(D \mid T_i)$ required for this purpose can be computed efficiently using Bayes' rule:

$$\Pr(D \mid T_i) = \Pr(T_i \mid D) \cdot \frac{\Pr(D)}{\Pr(T_i)}$$

The probability distribution $\Pr(D)$ and $\Pr(T_i)$ for each test variable T_i are already available after the last test result had been entered and therefore do not require any further propagations. The probability distributions $\Pr(T_i \mid D)$ for all test variables can be established by propagating the various possible values for the disease variable D. The number of propagations required to establish these distributions for all test variables simultaneously thus equals the number of values of D.

We have argued above that the test-selection strategy implied by this basic myopic algorithm seems to be an oversimplification of the test-selection routines that are found in many medical fields of application. To arrive at a test-selection facility that fits in more closely with daily practice, we enhance the basic myopic algorithm to take a sequence of different subgoals into consideration. To this end, the algorithm is extended to take a list $\mathcal{S}$ of subgoals S_i as part of its input. The algorithm now performs test selection per subgoal, that is, for each current subgoal it focuses on the test variables that provide information about that particular subgoal. For this purpose, the algorithm has to be provided with information about which variables provide information about which subgoals. This information can in essence be computed automatically from the probabilistic network under consideration, for example, using feature selection techniques [30,35]. However, since this type of information is rather easily given by domain experts, we decided to provide it as part of the input to the algorithm. For every subgoal S_i, all test variables that provide information about that specific subgoal, are included in the subset $\mathcal{T}(S_i)$ of $\mathcal{T}$.

We would like to note that the subsets $\mathcal{T}(S_i)$ do not have to be mutually disjoint: test variables may provide information about various subgoals.

In the presence of subgoals, we further need to reconsider the stopping criterion that is to be used by the algorithm. We recall that such a criterion is included to prevent overtesting. The stopping criterion should now take the subgoals into consideration as well as the overall goal. While selecting test variables for a specific subgoal, for example, the algorithm should be allowed to move to the next subgoal when no further information is needed with regard to the first subgoal even though there still are some tests left for that goal. Also, if for a specific subgoal the stopping criterion indicates that the test-selection process should continue, the algorithm should be allowed to stop selecting tests altogether if the value of the overall goal indicates that further testing is not necessary. The stopping criterion will be discussed in more detail in Section 7.2; for now it suffices to note that we will use two separate variables in our algorithms to provide for a stopping criterion that takes the subgoals and the overall goal into consideration.

Our first algorithm now computes the most informative test to be performed at the level of single test variables. The user is prompted for just the selected test variable and only the evidence for this variable is propagated throughout the network, before the test-selection process is continued.

Algorithm A_1: myopic test selection with subgoals

input:
$\mathcal{S}$: list of subgoals S_i
$\mathcal{T}$: list of test variables T_j, organised in sublists $\mathcal{T}(S_i)$ per subgoal S_i

Stop-subgoal(S_i)=*false*
Stop-overall=*false*
while $\mathcal{S} \neq \varnothing$ and *Stop-overall*≠*true* **do**
    select next S_i from $\mathcal{S}$
    remove S_i from $\mathcal{S}$
    while $\mathcal{T}(S_i) \neq \varnothing$, *Stop-subgoal*$(S_i)$≠*true* and *Stop-overall*≠*true* **do**
        compute most informative $T_j \in \mathcal{T}(S_i)$
        prompt for evidence for T_j and propagate
        remove T_j from $\mathcal{T}$
        compute *Stop-subgoal*(S_i)
        compute *Stop-overall*
    od
od

As an example of the concept of subgoal, we consider again the domain of oesophageal cancer. From our interviews with the experts, as described in Chapter 6, the various different subgoals in their test-selection strategy are readily established. These subgoals are the physical condition of the patient, the presence or absence of distant metastases, the resectability of the primary tumour, and the presence of contra-indications for surgery, which are to be addressed in the indicated order. For the subgoal pertaining to the presence of distant metastases, for example, the relevant

diagnostic tests are a CT scan of the abdomen, a CT scan of the thorax, an endosonography of the local region of the primary tumour and of the truncus coeliacus, a laparoscopic examination of the liver and truncus coeliacus, a physical examination of the neck, an ultrasound of the neck, and an X-ray of the patient's chest. All variables that serve to capture results of these tests are included in the set of test variables for that subgoal.

The algorithm now proceeds as follows. It selects the first subgoal from the list $\mathcal{S}$ of subgoals. It then selects the set of all test variables that provide information about the selected subgoal. From among these test variables, it selects the variable that is expected to yield the largest decrease in diagnostic uncertainty. The user then is prompted to enter evidence for the selected variable T_j. The entered evidence is propagated throughout the network, resulting in the posterior probability distribution over all variables given the entered evidence. Note that the evidence may have changed the probability distributions for the main diagnostic variable and for the variable representing the subgoal under study. The process of selecting test variables continues until the stopping criterion for the subgoal under study or that for the overall goal has been met, or all tests for the subgoal have been performed. When the stopping criterion for the subgoal is satisfied, or the set of test variables has been exhausted, the algorithm selects the next subgoal. As soon as the overall stopping criterion is satisfied, the test-selection process is halted.

Algorithm A_1 is strictly myopic in its test-selection strategy: test variables are selected sequentially on a one-by-one basis and the next test variable is selected only after the user has entered evidence for the previous one. We have argued above that a myopic test-selection strategy may not be realistic for many applications in medicine. We have also argued that a fully non-myopic strategy may be infeasible for practical purposes. Our second algorithm now is non-myopic in nature, yet uses a predefined clustering of the test variables. The clustering of the test variables in the network is given as part of the input to the algorithm. The algorithm equals

Algorithm A_2: non-myopic test selection with subgoals

input:
$\mathcal{S}$: list of subgoals S_i
$\mathcal{T}$: list of clusters C_j of test variables, organised in sublists $\mathcal{T}(S_i)$ per subgoal

Stop-subgoal(S_i)=*false*
Stop-overall=*false*
while $\mathcal{S} \neq \varnothing$ and *Stop-overall*≠*true* **do**
    select next S_i from $\mathcal{S}$
    remove S_i from $\mathcal{S}$
    while $\mathcal{T}(S_i) \neq \varnothing$, *Stop-subgoal*$(S_i)$≠*true* and *Stop-overall*≠*true* **do**
        compute most informative cluster $C_j \in \mathcal{T}(S_i)$
        prompt for evidence for all $T_k \in C_j$ and propagate
        remove C_j from $\mathcal{T}$
        compute *Stop-subgoal*(S_i)
        compute *Stop-overall*
    od
od

For the oesophageal cancer network, for example, we cluster the various different test variables according to the physical test they pertain to. The test variables that model the results of a gastroscopic examination with respect to the circumference, the length, the location, and the shape of the primary tumour, and the absence or presence of necrosis in the oesophagus, for instance, are included in a single cluster.

We would like to note that algorithm A_2 is much more computationally demanding than algorithm A_1. The increase of computation time stems from the computation of the most informative cluster of test variables. In Chapter 5 we argued that the expected decrease in the Gini index after gathering evidence for a single test variable is established by computing the Gini index given each possible result for the variable and then weighing these by the probabilities that the respective results occur. More formally, the expected Gini index $G(\Pr(D \mid T_j))$ after gathering evidence for the test variable T_j is defined as the expected value of the Gini index where the expectation is taken over all possible test results t_j^k:

$$G(\Pr(D \mid T_j)) = \sum_{k=1,\ldots,m_j} G(\Pr(D \mid T_j = t_j^k)) \cdot \Pr(T_j = t_j^k)$$

We now observe that in algorithm A_2, we need to compute the expected Gini index with respect to a cluster C_j of test variables. To this end, the Gini index $G(\Pr(D \mid C_j = c))$ has to be computed for each combination of values $c = t_1^{k_1}, \ldots, t_n^{k_n}$ of the test variables $T_1, \ldots T_n \in C_j$. In addition, the joint probability $\Pr(C_j = c)$ has to be established for each such combination. The probability distribution $\Pr(D \mid C_j = c)$ for a specific combination of values c can be computed by a single propagation throughout the network. The computation of all distributions $\Pr(D \mid C_j = c)$ now requires at most 2^n propagations for n binary test variables; note that using the concept of d-separation, this number of propagations can often be reduced in practice. The computation of each separate probability $\Pr(C_j = c)$ already requires n propagations, building upon the chain rule $\Pr(t_1^{k_1}, \ldots, t_n^{k_n}) = \Pr(t_1^{k_1} \mid t_2^k, \ldots, t_n^k) \cdot \ldots \cdot \Pr(t_{n-1}^{k_{n-1}} \mid t_n^{k_n}) \cdot \Pr(t_n^{k_n})$. The computation of the entire joint distribution $\Pr(C_j)$ thus requires at most $n \cdot 2^n$ propagations. Note that the computation of the expected Gini index with respect to clusters of test variables takes an exponential number of propagations, while the computation of the expected index with respect to all variables separately takes just a constant number of propagations. Further note that algorithm A_2 rapidly becomes infeasible for practical purposes as the number of test variables per cluster increases.

Algorithm A_2 in essence is non-myopic in its test-selection strategy. As we argued above, this non-myopic algorithm might become computationally too demanding, especially if a meaningful clustering of test variables results in clusters of relatively large size. A blood test, for instance, may easily yield results for tens of variables in a network. To save computation time yet retain some of the idea of non-myopia, we designed an algorithm that implies a *semi-myopic* test-selection strategy. This semi-myopic algorithm equals

Algorithm A_3: semi-myopic test selection with subgoals

input:
$\mathcal{S}$: list of subgoals S_i

$\mathcal{T}$: list of clusters C_j of test variables, organised in sublists $\mathcal{T}(S_i)$ per subgoal

Stop-subgoal(S_i)=*false*
Stop-overall=*false*
while $\mathcal{S} \neq \varnothing$ and *Stop-overall*≠*true* **do**
    select S_i from $\mathcal{S}$
    remove S_i from $\mathcal{S}$
    while $\mathcal{T}(S_i) \neq \varnothing$, *Stop-subgoal*$(S_i) \neq$ *true* and *Stop-overall* $\neq$ *true* **do**
        compute most informative $T_j \in \mathcal{T}(S_i)$
        prompt for evidence for T_j and for all $T_k \in C_m$ with C_m such that $T_j \in C_m$,
        and propagate
        remove C_m from $\mathcal{T}$
        compute *Stop-subgoal*(S_i)
        compute *Stop-overall*
    od
od

The algorithm very much resembles the myopic algorithm A_1 presented above. The main difference is that algorithm A_3 prompts not just for the result of the selected test variable T_j, but for the results of all test variables T_k that belong to the same cluster as the selected variable. As an example, we consider again the domain of oesophageal cancer. Suppose that from the network the test variable *CT-liver* is selected as the most informative variable in some step of the test-selection process. The user now is prompted by the algorithm for the value of this variable. The user is also prompted for a value for the variable *CT-truncus*, since this test variable models a result that is obtained from the same physical test as the variable *CT-liver*. Entering evidence for physical tests rather than for just one test variable seems to fit in more closely with the daily practice of the physicians, since in daily practice they are inclined to think in terms of physical tests even though they may be interested mainly in the value of a single variable. After performing the test, therefore, it seems logical to enter not only the result that is currently of interest, but all other results obtained from the same physical test as well.

From the results of our test-selection experiments described in Chapter 5, we observed that the test-selection sequences do not differ that much among the patient records for which we ran the experiment. This appears to imply that specific test variables are highly informative for many patients, while other variables provide less valuable information. For instance, the CT-scan of the patient's liver was indicated as the first most informative test result for a large number of patients. Since some test variables are highly informative, the physical test they pertain to is likely to be quite informative as well. We thus do not expect large differences in the test-selection sequences resulting from the three different algorithms. Note that the three algorithms may in general behave rather differently.

7.2 Stopping criteria

In the previous section, we presented three algorithms for test selection that have been designed to be better tailored to medical practice than the basic myopic test-selection facility commonly in use. The new algorithms include a stopping criterion to prevent overtesting. Since the algorithms aim at test selection with regard to a sequence of subgoals, in addition to an overall goal, we now have to design our criterion to apply to different goals. We recall from Chapter 3 that the most commonly used stopping criterion builds upon the probability of the most likely value of the main diagnostic variable. We propose to apply this criterion to the various subgoals distinguished. We further observe that, since we use the expected decrease of the value of the Gini index over a specific goal variable as the basis for test selection, this expected decrease can be taken into consideration for a stopping criterion as well. We now propose four different criteria.

Criterion 1
The first stopping criterion examines whether or not there exists a value of the goal variable under study, that has a probability equal to or greater than a prespecified probability threshold. The basic idea of this stopping criterion is that, if a specific value of the goal variable has a probability greater than, for example, 0.90, then the attending physician will be confident enough to act as if this value were the underlying truth. More formally, we consider a goal variable D with values $d_j, j = 1, \ldots, m$, and a threshold value $0 < \alpha < 1$. The selection of further tests now is halted as soon as

$$\exists p \in \{\Pr(D = d_j) \mid j = 1, \ldots, m\} : p \geq \alpha$$

where $\Pr$ denotes the probability distribution over the goal variable given all previously entered results. Note that the distribution over the goal variable is examined prior to considering ordering another test. The possible influences that the result of a subsequent test can have on the distribution, therefore are not taken into account. The larger the value of α chosen, the longer the resulting test-selection facility will continue to select tests. Note that if all tests in a domain of application have low sensitivities and specificities, as described in Chapter 4, it is not very likely that a sequence of test results will give rise to a high probability of the most likely value. The lower the sensitivities and specificities, therefore, the lower the value of α should be in order to prevent overtesting. From a computational point of view, we would like to note that the criterion does not require any additional propagations, since the distribution $\Pr$ to be examined is available after the last evidence has been propagated.

Criterion 2
Our second stopping criterion, like the first one, also considers the probability distribution over the goal variable. While the first stopping criterion focused on the most likely value of this variable, however, the second criterion reviews the entire probability distribution over all values. The criterion thereby takes the skewedness of the distribution into consideration. As an example, we consider a goal variable D with m possible values, where d_j is the most likely one. The remaining probability mass can for example be distributed evenly over the remaining $m - 1$

values for D or it can be focused on a single value $d_i, i \neq j$. Taking the entire distribution into consideration, we may wish to continue testing in the latter case and to abstain from further tests in the former case. Multiple measures have been designed that serve to review a complete probability distribution, among which is the Gini index. Since we already use the Gini index in our test-selection algorithms for measuring uncertainty, we decided to use this measure for measuring the skewedness of the distribution as well. More formally, for a prespecified threshold value β, the selection of further tests is halted as soon as

$$G(\Pr(D)) \leq \beta$$

where $\Pr$ again denotes the probability distribution over the goal variable given all previously entered test results. Note that the choice of a value for β depends on the number of values of the goal variable D. For instance, if D is a binary variable, β has to be chosen from between 0 and 0.5, which are the minimum and maximum value, respectively, of $G(\Pr(D))$. The smaller the value of β chosen, the longer the resulting test-selection facility may continue to select tests. From a computational point of view, we would like to note that the criterion does not require any additional propagations, since the probability distribution $\Pr$ to be examined is available after the last evidence has been propagated. Furthermore, the Gini index of this distribution is also already available since it has been computed when establishing the expected value of the Gini index given the test variable at hand.

Criterion 3
Our third stopping criterion, unlike the first two criteria, does not review the current probability distribution over the goal variable. The criterion is based upon a *one-step look-ahead* and considers the change that can be occasioned in the Gini index of the current distribution if we were to establish the value of one more test variable. As long as considerable changes are expected, the selection of test variables is continued. More formally, the criterion considers the expected decrease $\widetilde{G}(T_i)$ in the Gini index of the current distribution over the goal variable for each remaining test variable and examines whether or not this expected decrease for all test variables T_i is smaller than a prespecified threshold value δ. The selection of further tests then is halted as soon as

$$\widetilde{G}(T_i) \leq \delta$$

Note that the choice of a value for δ again depends on the number of values of the goal variable. Furthermore, the smaller the value of δ chosen, the longer the resulting test-selection facility may continue to select tests. From a computational point of view, we would like to note that evaluating the criterion requires a constant number of propagations, since the expected decreases in the Gini index after performing a single next test can be computed with a constant number of propagations using Bayes' rule as described above. Note that the now computed expected decreases in the Gini index of the probability distribution over the goal variable can be used in the next step of the test-selection process. Only in the step for which the stopping criterion indicates that the selection of tests should halt, does the criterion incur a constant number of extra propagations.

Criterion 4
Just like the third criterion, our fourth stopping criterion is based upon a one-step look-ahead. While the third criterion focused on the expected decrease in the value of the Gini index after performing a next test, our fourth criterion studies the decrease that can *maximally* be occasioned by a next test result. We note that in computing the expected decrease, the separate decreases from the different test results are weighted by the probabilities that these results will occur. It is possible, therefore, that a single test result occasions a large decrease in diagnostic uncertainty while the test variable itself is not expected to result in a substantial decrease. Our fourth stopping criterion now considers the decrease in the value of the Gini index of the probability distribution over the subgoal D that is occasioned by each result t_i^k of every test variable T_i separately. Our criterion examines whether or not all possible test results t_i^k of all test variables T_i will result in a decrease in the value of the Gini index of the probability distribution over D that is smaller than a prespecified threshold value ϵ. The selection of further tests is halted as soon as

$$\forall g \in \{G(\Pr(D \mid T_i = t_i^k)) \mid i = 1, \ldots, n, k = 1, \ldots, m_i\} : g \leq \epsilon$$

where $\Pr$ once again denotes the probability distribution over the goal variable given all previously entered test results in addition to the test result t_i^k and $T_1, \ldots, T_n$ are the test variables that pertain to the current subgoal. From a computational point of view, we would like to note that the criterion again requires a constant number of propagations, since the expected decreases in the Gini index after performing a single next test can be computed with a constant number of propagations using Bayes' rule as described above. Note that, again, the now computed expected decreases in the Gini index of the probability distribution over the goal variable can be used in the next step of the test-selection process. Again, only in the step for which the stopping criterion indicates that the selection of tests should halt, does the criterion incur a constant number of extra propagations. Note that the smaller the value of ϵ chosen, the longer the resulting test-selection facility may continue to select tests. We further would like to note that if our last stopping criterion is met, then our third criterion is met as well. The reverse property does not hold, however.

To conclude, we observe that our third and fourth criterion employ a one-step look-ahead. In essence, it is possible to construct a more involved stopping criterion by including an n-step look-ahead, $n > 1$. It will be evident, however, that the larger n, the harder it becomes from a computational point of view to evaluate such a criterion. We argued above that a stopping criterion with a one-step look-ahead requires a constant number of propagations to be performed. We further argued that the probabilities that are computed upon evaluating the criterion, can be re-used in the next step in the test-selection process. When using an n-step look-ahead, however, the computed probabilities can no longer be re-used. Further note that the more steps the criterion looks ahead, the more the resulting test-selection will start to resemble a full decision analysis.

7.3 Experiment

To evaluate and compare the performance of the three algorithms for test selection and the four stopping criteria described above, we conducted an experimental study in the context of the oe-

sophageal cancer network. Using this network we studied the behaviour of the four stopping criteria and addressed the question whether a particular criterion would show a tendency to stop too early in the test-selection process. We compared the sequences of test variables selected by the different algorithms using the second stopping criterion. We briefly evaluated the overall test-selection strategy implied by the myopic and semi-myopic algorithms with our two domain experts. The semi-myopic test-selection algorithm proved to result in the most preferred sequence of test variables.

7.3.1 The extended network

In our experimental study we used an extended version of the original oesophageal cancer network. We recall that the original network aims at establishing the stage of a patient's cancer; we further recall that the network contains 42 variables of which 25 variables serve to represent the various different test results. We now observe that, while the network includes a main diagnostic variable modelling the stage of a patient's cancer, none of its 42 variables capture the subgoals that we elicited during the interviews with our experts. Since our algorithms were designed to perform the selection of test variables with respect to these goals, we decided to extend the original oesophageal cancer network with a number of goal variables.

The newly included variables in essence model the five goals for test selection that are captured by the decision moments in the flowchart from Figure 6.5. The goal variable *Condition* describes the physical condition of the patient and has the two possible values *good* and *poor*. The goal variable *M1*, with the values *yes* and *no*, models the presence or absence of distant metastases. The variable *Resectable*, with the values *yes* and *no*, provides information about the resectability of the patient's primary tumour. The variable *Contra-indic-surgery* with the values *yes* and *no*, summarises whether or not there are indications for the patient not to undergo surgical removal of the oesophagus. The variable *Therapy*, to conclude, is the overall goal variable that captures three different treatment categories: *low-dose radiotherapy or endoprosthesis*, *high-dose radiotherapy possibly with chemotherapy*, and *surgery*. In addition to the five goals that we modelled explicitly, the test-selection process involves yet another goal variable *General information* for which we decided not to include an additional variable in our network. Since an attending physician will always order a gastroscopic examination combined with a biopsy, this goal was not included as a decision moment in the flowchart that captures the expert's test-selection strategy. Note that the goal basically serves to initiate the test-selection process and is not used itself in any further decision. During our interviews, we elicited for each subgoal which decision should be taken based upon the various different possible test results. The new goal variables are now included in the oesophageal cancer network by adding arcs from the variables that capture the underlying true conditions that influence the decisions. Note that by doing so, the sensitivity and specificity of the results are taken into account. The probability tables for the new variables only contain the values 0 and 1, thereby expressing that these variables themselves do not involve any uncertainty.

To conclude, we added to the original oesophageal cancer network also the variables *Smoking*, *WHO-status*, and *Age* since the values of these variables are needed for establishing a patient's condition and to decide upon the presence or absence of contra-indications for surgery.

The extended network is shown in Figure 7.1.

7.3.2 Information about test variables and goals

As described in Section 7.1, our new test-selection algorithms need extra information for their input in addition to the probabilistic network under consideration. The extra information in essence consists of three lists: a list of the subgoals for the test-selection process, a list that describes for each subgoal which test variables provide information about that particular subgoal, and a list for each physical test that describes which test variables serve to capture the results of that test.

For the various different subgoals and the order in which they are to be examined, we build in our experiment upon the sequence that was elicited from our domain experts and captured in the flowchart of Figure 6.5. The first subgoal to be addressed is the implicit *General information* goal; the second goal is captured by the variable *Condition*. We assume that every patient has to undergo a gastroscopy with a biopsy and has a consult with his physician. The diagnostic tests that serve to provide information about these two goals should therefore always be ordered and no selection of tests will take place. We will call such goals an *easy goal*. After the physical condition of the patient has been established, the next goal is *M1*, which is to determine the presence or absence of metastases distant from the primary tumour. After this goal has been addressed, the subgoal modelled by the variable *Resectability* is the next one under study. The last subgoal equals the presence or absence of contra indications for surgery. For the list of subgoals $\mathcal{S}$ to be provided as input to our test-selection algorithms, we thus have that $\mathcal{S}$ = (*General information*, *Condition*, *M1*, *Resectability*, *Contra-indic-surgery*). Note that the overall goal, captured by the variable *Therapy*, is not included in the sequence and therefore is never addressed explicitly in the test-selection process. We recall that this overall goal was included into the network to provide for an overall stopping criterion.

In addition to the list of subgoals, our algorithms require for their input information about which test variables pertain to which subgoals. From the interviews that we had with our experts and the resulting flowchart, we readily determined the required information: Table 7.1 summarises the relationship between the test variables on the one hand and the goals on the other hand. From the table, we have that the age of the patient, his smoking habits and his WHO-status directly provide information about the presence or absence of contra indications for surgery. Further tests required for this purpose, which include blood samples and an ECG, unfortunately have never been modelled in the original network; also, our data collection does not include any results for these tests.

The third list to be provided as part of the input to our algorithms, pertains to the relationship between test variables on the one hand and physical tests on the other hand. The 28 test variables included in the extended oesophageal cancer network describe the results of 12 different physical tests. These physical diagnostic tests are an *interview*, a *biopsy*, a *bronchoscopy*, a *CT-scan of the abdomen*, a *CT-scan of the thorax*, an *endosonography of the oesophagus*, a *sonography* of the neck, a *gastroscopy*, a *laparoscopy*, a *physical examination*, a *barium swallow*, and an *X-ray* of the patient's chest. Table 7.2 matches the various test variables in the network to these physical tests.

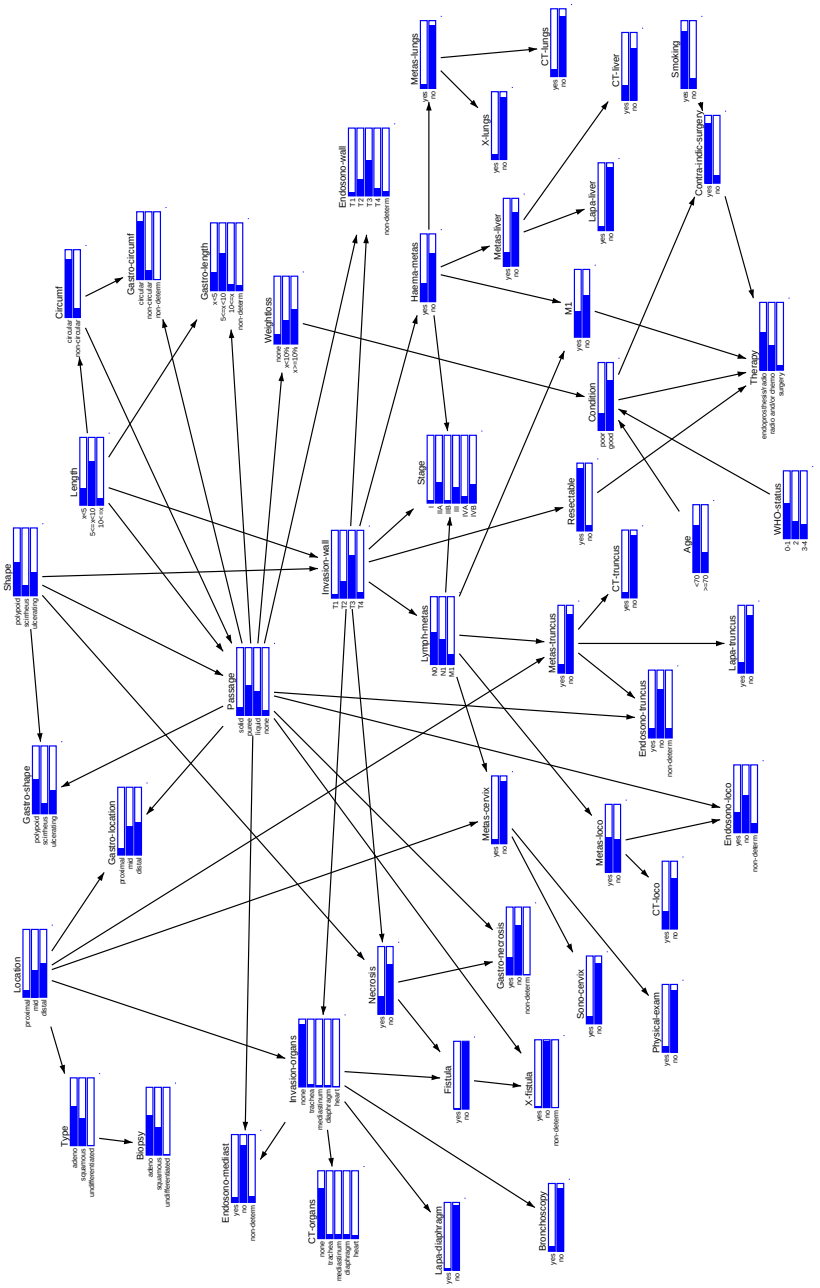


Figure 7.1: The extended oesophageal cancer network.

Table 7.1: An overview of the test variables in the extended oesophageal cancer network and the subgoal(s) they pertain to.

variable	**goal**
Age	Condition, Contra-indic-surgery
Barium swallow	Therapy
Biopsy	General information
Bronchoscopy	Resectability
CT-liver	M1
CT-loco	M1
CT-lungs	M1
CT-organs	Resectability
CT-truncus coeliacus	M1
Endo-sonography loco	M1
Endo-sonography mediastinum	Resectability
Endo-sonography truncus coeliacus	M1
Endo-sonography oesophageal wall	Resectability
Gastroscopy circumference	General information
Gastroscopy length	General information
Gastroscopy location	General information
Gastroscopy necrosis	General information
Gastroscopy shape	General information
Laparoscopy liver	M1
Laparoscopy diaphragm	Resectability
Laparoscopy truncus coeliacus	M1
Passage	General information
Physical examination neck	M1
Smoking	Condition, Contra-indic-surgery
Sonography neck	M1
Weightloss	Condition
WHO status	Condition, Contra-indic-surgery
X-ray lungs	M1

Table 7.2: An overview of the test variables in the extended oesophageal cancer network and the physical test they pertain to.

variable	**diagnostic test**
Age	Interview
Barium swallow	Barium swallow
Biopsy	Biopsy
Bronchoscopy	Bronchoscopy
CT-liver	CT-abdomen
CT-loco	CT-thorax
CT-lungs	CT-thorax
CT-organs	CT-thorax
CT-truncus coeliacus	CT-abdomen
Endo-sonography loco	Endosonography
Endo-sonography mediastinum	Endosonography
Endo-sonography truncus coeliacus	Endosonography
Endo-sonography oesophageal wall	Endosonography
Gastroscopy circumference	Gastroscopy
Gastroscopy length	Gastroscopy
Gastroscopy location	Gastroscopy
Gastroscopy necrosis	Gastroscopy
Gastroscopy shape	Gastroscopy
Laparoscopy liver	Laparoscopy
Laparoscopy diaphragm	Laparoscopy
Laparoscopy truncus coeliacus	Laparoscopy
Passage	Interview
Physical examination neck	Physical examination
Smoking	Interview
Sonography neck	Sonography
Weightloss	Interview
WHO status	Interview
X-ray lungs	X-ray

7.3.3 The stopping criteria evaluated

We implemented the four stopping criteria discussed in Section 7.2 and studied their performance with algorithm A_1 for myopic test selection with subgoals. We decided to study the criteria in relation with algorithm A_1 since this algorithm allows for studying a larger variety of test-selection steps than the algorithms A_2 and A_3. We recall that algorithm A_1 considers each test variable separately, while the other two algorithms prompt the user for the results of physical tests and, hence, for a number of test variables at a time. Since our network contains just 12 physical tests, 9 of which are involved in the actual test selection process, using one of the other algorithms would not allow for studying the behaviour of the different criteria in detail. Note, in contrast, that our network includes 28 test variables, 18 of which are involved in the test selection process. In our experiment, we further focused on the subgoal captured by the variable *M1*. We decided to focus on the subgoal *M1* since, from among the various subgoals, the largest number of test variables provide information about this goal. We thus could focus on a sufficiently large set of test variables to examine the behaviour of the criteria. Before the selection of tests is initiated for the subgoal *M1*, we entered, for each patient, the results of the tests pertaining to the two easy goals, *General information* and *Condition*. We ran the experiments with the eight patient cases that we had designed for the second interview with our domain experts as described in Chapter 6. Since these cases had been designed to lie at the decision boundaries of our experts' test-selection strategy, we expect the various different stopping criteria to behave representatively with respect to the data set of real patients. Note that we started the evaluation of the stopping criteria with respect to the subgoal *M1*. Patients who are in a poor physical condition therefore still receive the basic package of tests.

For the first patient, the following test results are available that provide information about the first two easy subgoals:

- *Age* ≥ 70 *years*
- *Biopsy* = *squamous*
- *Gastro-circumf* = *non-circular*
- *Gastro-length* ≥ 10 *cm*
- *Gastro-location* = *mid*
- *Gastro-necrosis* = *no*
- *Gastro-shape* = *scirrheus*
- *Passage* = *solid*
- *WHO-status* = 2
- *Weightloss* $\geq 10\%$

From these results, we have that the condition of the patient is poor. The experts' strategy now indicates that the test-selection process should halt. However, we now assume the patient's condition to be good and continue the test-selection process with the subgoal *M1*. The first test variable that is selected by our algorithm with respect to this goal, is the variable *CT-liver*. The result for this test variable is *no* and this value is entered into the network. The result indicates that, upon performing the CT-scan, no metastases have been found in the patient's liver. The test-selection algorithm should not decide to abstain from further testing, however, since not all sites distant from the primary tumour have yet been examined. The second test variable selected by our algorithm is the variable *Lapa-truncus* which also has for its result the value *no*. Again, the test-selection algorithm should decide to continue: metastases could, for example, still be located in the patient's lungs which have not been examined as yet. The third selected test variable is the variable *X-lungs*. The result of the X-ray of the patient's lungs is *yes*, indicating that metastases are present in the patient's lungs. Since evidence of metastases distant from the primary tumour has been found, the test-selection process may now be halted. Since the sensitivity and specificity of the X-ray are not very high, we feel that performing an additional CT-scan of the lungs would be acceptable, however. The entire sequence of test variables that would be selected for the subgoal *M1* if no stopping criterion would be employed is

- *CT-liver = no*
- *Lapa-truncus = no*
- *X-lungs = yes*
- *CT-lungs = yes*
- *Lapa-liver = no*
- *CT-truncus = no*
- *Endo-sono-loco = no*
- *CT-loco = yes*
- *Sono-cervix = yes*
- *Physical-exam = no*
- *Endo-sono-truncus = no*

From the above analysis, we conclude that, for the first patient, the test-selection process should stop after the result of the third, or maybe fourth, test variable has been entered and propagated throughout the network. Similar analyses were made of the sequences of diagnostic tests selected for the other patients. We found that, for the patients 3, 4, 5 and 7, the test-selection process should not stop at all; for patient 2, the selection of tests can stop after the result for the eighth variable has been entered; for patient 6 the test-selection process can be halted after the third test has been entered and propagated throughout the network and for patient 8 after the result for the

Table 7.3: Criterion 1: the interval for which the stopping criterion was met for the patients 1, 2, 3 and 4.

α	matched interval patient 1	matched interval patient 2	matched interval patient 3	matched interval patient 4
0.975	[. . .]	[. . .]	[. . .]	[. . .]
0.95	[5 . . . 11]	[. . .]	[. . .]	[. . .]
0.925	[5 . . . 11]	[. . .]	[. . .]	[. . .]
0.9	[5 . . . 11]	[. . .]	[. . .]	[4 . . . 6]
0.875	[5 . . . 11]	[. . .]	3,4, 10	[3 . . . 11]
0.85	[5 . . . 11]	[. . .]	3,4 [6 . . . 11]	[3 . . . 11]
0.825	3, [5 . . . 11]	[5 . . . 7]	[3 . . . 11]	[3 . . . 11]

Table 7.4: Criterion 1: the interval for which the stopping criterion was met for the patients 5, 6, 7 and 8

α	matched interval patient 5	matched interval patient 6	matched interval patient 7	matched interval patient 8
0.975	[. . .]	[5 . . . 11]	[. . .]	[5 . . . 11]
0.95	[. . .]	[4 . . . 11]	[. . .]	[4 . . . 11]
0.925	[. . .]	[4 . . . 11]	7	[4 . . . 11]
0.9	[. . .]	[4 . . . 11]	6,7	[4 . . . 11]
0.875	[. . .]	[4 . . . 11]	[4 . . . 7]	[4 . . . 11]
0.85	5,6,9	[4 . . . 11]	[3 . . . 9]	[4 . . . 11]
0.825	[3 . . . 11]	[4 . . . 11]	[3 . . . 11]	[4 . . . 11]

second test variable has been entered. Note that, although the flowchart indicates for patients 1 and 7 that no further testing is required because of their physical condition, we decided to nevertheless perform test selection with respect to the subgoal *M1*, to allow for studying the behaviour of the various stopping criteria for a wide range of patient profiles. We would further like to note that in the medical domain, stopping too early is considered worse than stopping too late. When the process of gathering evidence is halted too early, incorrect conclusions can be made. For the first patient discussed above, for example, if the test-selection process would be halted after a value for the second variable had been entered, the erroneous conclusion that the patient does not have any distant metastases, would be drawn. From the point of view of diagnostic information, stopping too late is not an important issue. Performing too many tests nonetheless has consequences for the discomfort of the patient and the amount of money spent for example.

Tables 7.3-10 report the results obtained with the four stopping criteria employing different threshold values, for the eight patients under study. As an example, we consider the first column of Table 7.3, which summarises the intervals for which the first stopping criterion was satisfied for the first patient, in closer detail. With $\alpha = 0.975$, stopping criterion 1 was never met for our patient, which is represented in the table as $[\ldots]$. With $\alpha = 0.975$ therefore, the test-selection

Table 7.5: Criterion 2: the interval for which the stopping criterion was met for the patients 1, 2, 3 and 4.

β	matched interval patient 1	matched interval patient 2	matched interval patient 3	matched interval patient 4
0.05	[. . .]	[. . .]	[. . .]	[. . .]
0.075	[5 . . . 11]	[. . .]	[. . .]	[. . .]
0.1	[5 . . . 11]	[. . .]	[. . .]	[. . .]
0.125	[5 . . . 11]	[. . .]	[. . .]	[. . .]
0.15	[5 . . . 11]	[. . .]	[. . .]	[. . .]
0.175	[5 . . . 11]	[. . .]	[. . .]	[4 . . . 6]
0.2	[5 . . . 11]	[. . .]	4,10	[3 . . . 6]

Table 7.6: Criterion 2: the interval for which the stopping criterion was met for the patients 5, 6, 7 and 8.

β	matched interval patient 5	matched interval patient 6	matched interval patient 7	matched interval patient 8
0.05	[. . .]	[5 . . . 11]	[. . .]	[5 . . . 11]
0.075	[. . .]	[5 . . . 11]	[. . .]	[4 . . . 11]
0.1	[. . .]	[4 . . . 11]	[. . .]	[4 . . . 11]
0.125	[. . .]	[4 . . . 11]	[. . .]	[4 . . . 11]
0.15	[. . .]	[4 . . . 11]	6,7	[4 . . . 11]
0.175	[. . .]	[4 . . . 11]	6,7	[4 . . . 11]
0.2	[. . .]	[4 . . . 11]	[4 . . . 7]	[4 . . . 11]

Table 7.7: Criterion 3: the interval for which the stopping criterion was met for the patients 1, 2, 3 and 4.

δ	matched interval patient 1	matched interval patient 2	matched interval patient 3	matched interval patient 4
0.01	[5 . . . 11]	[4 . . . 11]	4, [6 . . . 11]	[4 . . . 11]
0.02	3, [5 . . . 11]	[4 . . . 11]	[3 . . . 11]	[3 . . . 11]
0.03	3, [5 . . . 11]	[3 . . . 11]	[3 . . . 11]	[3 . . . 11]
0.04	3, [5 . . . 11]	[3 . . . 11]	[3 . . . 11]	[3 . . . 11]
0.05	3, [5 . . . 11]	[3 . . . 11]	[3 . . . 11]	[3 . . . 11]
0.06	3, [5 . . . 11]	[3 . . . 11]	[3 . . . 11]	[3 . . . 11]
0.07	2,3, [5 . . . 11]	[3 . . . 11]	[2 . . . 11]	[2 . . . 11]

Table 7.8: Criterion 3: the interval for which the stopping criterion was met for the patients 5, 6, 7 and 8.

δ	matched interval patient 5	matched interval patient 6	matched interval patient 7	matched interval patient 8
0.01	[4 . . . 11]	[4 . . . 11]	[4 . . . 11]	[4 . . . 11]
0.02	[4 . . . 11]	[4 . . . 11]	[3 . . . 11]	[4 . . . 11]
0.03	[3 . . . 11]	[4 . . . 11]	[3 . . . 11]	[4 . . . 11]
0.04	[3 . . . 11]	[4 . . . 11]	[3 . . . 11]	[4 . . . 11]
0.05	[3 . . . 11]	[4 . . . 11]	[3 . . . 11]	[3 . . . 11]
0.06	[3 . . . 11]	[4 . . . 11]	[3 . . . 11]	[3 . . . 11]
0.07	[3 . . . 11]	[3 . . . 11]	[2 . . . 11]	[3 . . . 11]

Table 7.9: Criterion 4: the interval for which the stopping criterion was met for the patients 1, 2, 3 and 4.

ϵ	matched interval patient 1	matched interval patient 2	matched interval patient 3	matched interval patient 4
0.01	7,8,11	11	11	[. . .]
0.02	[7 . . . 11]	11	11	10
0.03	[6 . . . 11]	11	[9 . . . 11]	10,11
0.04	[6 . . . 11]	11	[9 . . . 11]	10,11
0.05	[5 . . . 11]	7,10,11	[9 . . . 11]	10,11
0.06	[5 . . . 11]	7,10,11	[9 . . . 11]	10,11
0.07	[5 . . . 11]	7,10,11	[9 . . . 11]	10,11

Table 7.10: Criterion 4: the interval for which the stopping criterion was met for the patients 5, 6, 7 and 8.

ϵ	matched interval patient 5	matched interval patient 6	matched interval patient 7	matched interval patient 8
0.01	[9 . . . 11]	[5 . . . 11]	[. . .]	[5 . . . 11]
0.02	[9 . . . 11]	[5 . . . 11]	11	[5 . . . 11]
0.03	[9 . . . 11]	[5 . . . 11]	11	[5 . . . 11]
0.04	6, [9 . . . 11]	[5 . . . 11]	9,11	[5 . . . 11]
0.05	6, [9 . . . 11]	[5 . . . 11]	9,11	[5 . . . 11]
0.06	6, [9 . . . 11]	[5 . . . 11]	9,11	[5 . . . 11]
0.07	6, [9 . . . 11]	[5 . . . 11]	[9 . . . 11]	[4 . . . 11]

process would not halt and our patient would have to undergo the entire sequence of diagnostic tests for the goal *M1*. For $\alpha = 0.9$, the interval $[5 \ldots 11]$ reported in the table indicates that the stopping criterion was met in all test-selection steps after a result for the fourth test variable had been entered. With $\alpha = 0.9$, therefore, our patient would only receive the first four tests, after which testing is stopped. With $\alpha = 0.825$, we observe that the stopping criterion was met before the third test variable was selected. With this threshold value, therefore, the test-selection process would have halted after just the first two tests had been performed. If the result for the third variable would have been entered, the stopping criterion would no longer have been met. Then, after the result for the fourth variable would have been entered, the criterion would again have been met. The results for $\alpha = 0.825$ are summarised in the table as $3, [5 \ldots 11]$. As we noted before, if the test-selection process would have been halted after the second test had been performed, the system would have concluded that no distant metastases were present; this conclusion would have been premature and, in fact, incorrect.

Tables 7.3 and 7.4 report the results obtained with the stopping criterion that builds upon the probability of the most likely value of the goal variable. We observe that, with $\alpha = 0.975$, the test-selection process is always stopped too late or not at all. Note that for even larger values for α, this trend of stopping too late would continue. With $\alpha \leq 0.825$, the stopping criterion is almost always satisfied too early. Note that for even smaller values for α this trend of stopping too early will continue. With $\alpha = 0.925$, for example, the test-selection process was halted too late for two patients (25%). For one patient (13%), the process halted too early. For the remaining 5 patients (63%), the selection of tests stopped when we would expect it to stop based upon our knowledge of the domain. With $\alpha = 0.95$ the test-selection process was halted too late for two patients (25%), for the remaining 6 patients (75%), the selection of tests stopped when we would expect it to stop.

Tables 7.5 and 7.6 report the results obtained with the stopping criterion that builds upon the Gini index. We observe that, with $\beta = 0.05$, the algorithm always stopped too late or not at all. Note that for even smaller values for β this trend of stopping too late would continue. With $\beta = 0.2$, the stopping criterion is mostly satisfied too early or in the step in the process in which we would expect it to be satisfied: for only one patient, the selection of tests was halted too late. Note that for even larger values for β this trend of stopping too late would continue. With $\beta = 0.15$, the stopping criterion was met too early for just one patient (13%). For two patients (25%), the process stopped too late. For the remaining 5 patients (63%), the stopping criterion was met when we would wish it to be satisfied. With $\beta = 0.125$, the stopping criterion was met too late for two patients (25%). For the remaining 6 patients (75%), the stopping criterion was met when we would wish it to be satisfied based upon our knowledge of the domain.

We now focus on the third criterion, which is based upon the expected decrease in the value of the Gini index. Tables 7.7 and 7.8 report the results obtained with this criterion. Even with the small threshold value of $\delta = 0.01$, the test-selection process halted too early for five patients (63%). For one patient (13%), the criterion was met too late. For the remaining two patients (25%), the test-selection process halted when we would wish it to halt. With $\delta = 0.04$, for example, the test-selection process halted too early for six patients (75%), thereby demonstrating that for larger values of δ, the trend of stopping too early continues. For one patient (13%), the criterion was met too late; for one patient (13%), the test-selection process halted when we would

wish it to halt.

Tables 7.9 and 7.10, to conclude, report the results obtained with the stopping criterion that builds upon the Gini index of all possible values of each test variable. We observe that with $\epsilon = 0.01$, the test-selection process halted too early for two patients (25%). For four patients (50%), the criterion was met too late. For the remaining two patients (25%), the test-selection process halted when we would wish it to halt. With $\epsilon = 0.04$, the test-selection process halted too early for four patients (50%). For four patients (50%), the criterion was met too late.

From our experimental results, we observe that both stopping criterion 1 and stopping criterion 2 perform well within our test-selection algorithm. Both criteria showed that the larger the threshold value for α and the smaller the threshold value for β, respectively, the less likely it becomes that the test-selection process will be halted too soon. Performing too few tests could thus easily be prevented. Overtesting, on the other hand, can also be prevented by selecting specific values for α and β, respectively. Choosing appropriate values for α and β now amounts to carefully weighing the effects of stopping too early and stopping too late. We observe that for both criteria we were able to extract reasonably good values for the thresholds from our experimental results. We recall that the first stopping criterion is based solely upon the probability of the most likely value of the goal variable under study. As we argued before, if two values are very likely and the remaining probability mass is spread over all other values, we would prefer to continue testing. On the other hand, if one value is very likely and the remaining probability mass is spread evenly over all other values, we would prefer to stop testing. Since the first stopping criterion is not able to distinguish between these two situations, we would expect it to perform worse than the second criterion which takes the spread of the probability mass into account. Although our experimental results do not support this expectation and show that both stopping criteria perform well, we conclude that more in general, stopping criterion 2 is the more preferred criterion among these two criteria.

From our experimental results, we further conclude that the third and fourth stopping criteria, which build upon the expected value of the Gini index after performing the next test, in practice did not meet our expectations in the sense that even for very small threshold values, they often gave rise to stopping too early. Our findings can be explained from the following observations. We would like the test-selection process to stop when the diagnostic uncertainty has been decreased to a small value and we would expect that no test, or test result alternatively, can induce a rather large change in the value of the Gini index. For the third and fourth stopping criterion, this would imply that we have to choose small values for δ and ϵ. If we take a close look at the function of the Gini index as plotted in Figure 5.1(b), we observe that, when the diagnostic uncertainty is the largest, shifts in the probability distribution may result in very small shifts in the value of the Gini index. With small values for δ and ϵ, therefore, the test-selection facility will tend to halt the process of gathering information also for relatively uniform distributions. This explains why the third and fourth stopping criterion are likely to result in stopping the selection of tests too early. Stopping criteria 3 and 4 as a consequence are not to be preferred.

We conclude this section by introducing a slightly more involved stopping criterion that we will use for our experiments with the three algorithms. Our main motivation for introducing a more involved criterion is our previous observation that stopping too early is considered worse than stopping too late. The basic idea of the more involved criterion now is that our stopping cri-

terion 2 is extended with an additional provision to forestall stopping too early. As we explained above, we prefer to use the value of the Gini index of the probability distribution over the goal variable as the basis for our stopping criterion. Now, if the criterion is met, indicating that there is not much uncertainty left with regard to the most likely value, then it may very well be that a test result exists, for a not yet performed test, that serves to again increase the diagnostic uncertainty. If this result would become available, then the stopping criterion would no longer be met in the next step of the test-selection process. This situation occurs in our experiments for patient 3, for example, with $\beta = 0.20$. After the first three test results were entered for this patient, the stopping criterion was met and the test-selection process was halted. Note that it was halted too early. If the fourth result would have been entered, however, the criterion would no longer have been met, thereby indicating that the test-selection process would have to be continued. To forestall stopping too early in situations like these, we propose to extend our stopping criterion 2 with a limited look-ahead.

Our new stopping criterion can be looked upon as a combination of the basic criteria 2 and 4. With the new criterion, the test-selection algorithm first employs the basic stopping criterion 2 to decide whether it should simply continue selecting tests or not. Now suppose that after the result for a test variable has been entered and propagated, the value of the Gini index over the subgoal is smaller than or equal to the prespecified threshold value β. Instead of halting the test-selection process immediately, the algorithm now employs a limited look-ahead to decide whether or not the next most informative test should still be ordered. This look-ahead amounts to examining whether or not a next test result could induce a shift in the value of the Gini index to such an extent that it no longer falls below the threshold value β. Note that this look-ahead basically amounts to employing our basic criterion 4. Only if the look-ahead supports the decision to stop gathering further information, is the test-selection process actually halted. If the basic criterion 2 is satisfied and the look-ahead indicates that the test-selection process cannot be halted yet, the algorithm continues its selection of tests. Experiments showed for patient 4, for instance, that the test-selection process originally halted with $\beta = 0.175$ after the result of the third test had been entered and propagated. Using our more involved stopping criterion, we observed that the test-selection process did not halt at all, as wished for. For patient 7, for example, our more involved stopping criterion with $\beta = 0.175$ resulted in the selection of all available test variables, as wished for. We would like to note that our more involved stopping criterion, due to the use of two different criteria, thus is capable of bridging more than one step. Note that the look-ahead is used only when criterion 2 is already satisfied. It is used, therefore, only if the Gini index is already rather small. The problems with criterion 4 that we elaborated upon above now do not occur. To summarise, the value of the *Stop-subgoal* variable mentioned in our algorithms, is thus computed with the following function:

Function: stopping criterion

input:
S_i: current subgoal
$\mathcal{T}(S_i)$: list of test variables T_j for subgoal S_i.

Stop-subgoal$(S_i) =$ *false*

compute $G(\Pr(D))$
if $G(\Pr(D)) \leq \beta$
  while $\mathcal{T} \neq \varnothing$ and *Stop-overall*$\neq$*true* **do**
    select $T_j \in \mathcal{T}(S_i)$
    for all test results t_j^k of T_j **do**
      compute $G(\Pr(D \mid T_j = t_j^k)), k = 1, \ldots m_j$
    od
    if for all $g \in \{G(\Pr(D \mid T_j = t_j^k)) \mid k = 1, \ldots, m_j\} : g \leq \beta$
      Stop-subgoal(S_i) =*true*
    od
  od
  return *Stop-subgoal*(S_i)

We would like to note that for every subgoal S_i, in essence a different value for β can be selected, that is dependent upon the consequences of taking an incorrect decision with respect to that particular subgoal. In our experiments, however, we will use the same threshold value β for all subgoals.

7.3.4 The test-selection algorithms evaluated

We implemented the three test-selection algorithms described in Section 7.1 with the combined stopping criterion introduced above. Once again, we ran experiments for the eight patient cases that we had designed for our interviews with the experts. We were interested in whether or not the various algorithms would result in sequences of test variables that fit the test-selection strategy employed by our experts; we were aware, however, that in our experiments we did not yet take the costs involved into consideration and as a consequence we did not expect a perfect match. In addition, we were interested in whether or not the combined stopping criterion would indicate, with all algorithms, to stop testing at the exact step in the test-selection process the experts would want it to stop.

We recall that algorithm A_1 is fully myopic in nature, selecting the diagnostic tests to be performed based upon the informative values of the separate variables: the expected value of the Gini index of the probability distribution over the subgoal after performing a diagnostic test is computed for single test variables. After selecting the most informative test variable, the user is prompted only for the result of this variable. Algorithm A_3 also performs the selection of tests at the level of the separate variables. After selecting the most informative test variable, however, the user is prompted not only for the result of this variable, but for the results of *all* test variables for which a result is obtained from the same physical test as for the most informative test variable. In contrast with the algorithms A_1 and A_3, algorithm A_2 selects diagnostic tests at the level of physical tests: the expected value of the Gini index of the posterior probability distribution over the subgoal is computed given all possible combinations of results for the various different test variables that are clustered with respect to a single physical test. We would like to note that not all test variables for which a physical test will yield a result, pertain to the same subgoal. For computing the expected value of the Gini index after performing the test, we decided to take only those test variables into account that pertain to the subgoal under study. Upon prompting the user

for the test's results, however, the algorithm requests the results for *all* variables concerned, also for those that belong to a different subgoal, since the test has now been performed and its results are available.

For all eight patient cases, we ran the three different algorithms. We used the combined stopping criterion from the previous section, which is based upon the Gini index of the probability distribution over the subgoal and employs a limited one-step look-ahead; for this threshold value, we used $\beta = 0.125$. We would like to note that we employed the same stopping criterion with the same threshold value for all subgoals, including the implicit overall goal. Further note that we excluded the laparoscopic examination from our experiments, even though it serves to provide information about the subgoals *M1* and *Resectability*. We recall that prior to performing a laparoscopic examination, the physicians have to decide upon the absence or presence of contra-indications for surgery. Since we do not have any information about such contra-indications in the oesophageal cancer network, the results of our experiment would be negatively influenced if we were to consider the laparoscopy without taking the contra-indications into consideration.

We now discuss the results obtained for patient 2. For the patient under study, the following test results from the starting package of tests were available:

- *Age* < 70 *years*
- *Biopsy = squamous*
- *Gastro-circumference = circular*
- *Gastro-length* ≥ 10 *cm*
- *Gastro-location = distal*
- *Gastro-necrosis = yes*
- *Gastro-shape = polypoid*
- *Passage = liquid*
- *WHO-status = 2*
- *Weightloss* $< 10\%$

After these results were entered and propagated throughout the network, the Gini index of the probability distribution over the overall goal, *Therapy*, equalled 0.56. The stopping criterion therefore was not met for the overall goal and the next subgoal *M1* was selected by all three algorithms. We will now discuss the sequences of diagnostic tests or test variables selected by the separate algorithms. Algorithm A_1 selected the variable *CT-liver* as the most informative test variable with regard to the subgoal under study. The user was prompted for its results. The value *CT-liver=no* was entered and was subsequently propagated to establish the posterior distribution for the subgoal *M1* given the new information. The Gini index of this new distribution equalled 0.43. The stopping criterion was thus not met yet for subgoal *M1*; the same observation held

with respect to the overall goal. The algorithm therefore proceeded with selecting the most informative test variable in view of the current distribution over the subgoal *M1*. The now selected test variable was *CT-truncus*. The result *CT-truncus=no* was entered into the network and propagated. The Gini index of the updated distribution over *M1* now equalled 0.36. The stopping criterion was again not satisfied for the subgoal under study; the same held for the overall goal. The next test variable selected by the algorithm was *Endosono-truncus*. The result *Endosono-truncus=no* was entered into the network and propagated. The Gini index of the distribution over the subgoal now equalled 0.32. Again the stopping criterion was not met. Also the overall criterion was not satisfied. The next test variable for which the user was prompted for a result, was *X-lungs*. The result *X-lungs=yes* was entered into the network and propagated. The Gini index of the distribution over *M1* after performing this test was 0.45. Note that the uncertainty with regard to the most likely value of the subgoal *M1* increased as a consequence of the result of the X-ray that pointed to another value than all previously entered test results. Next, the result of the CT-scan of the lungs was prompted for. The result *CT-lungs=yes* was entered into the network. After propagation, the Gini index of the distribution over the subgoal equalled 0.06. The stopping criterion was now satisfied for the subgoal *M1*. The one-step look-ahead showed that no result for the remaining test variables could serve to increase the Gini index so as to rise above the threshold value. We note that the stopping criterion for the overall goal *Therapy* was met at this stage as well. The test-selection process was thus halted. The selection of test variables is summarised in the first column of Table 7.11.

Algorithm A_2 selected the CT scan of the patient's upper abdomen as the most informative physical test with regard to the subgoal under study. The user was prompted for its results. The values *CT-liver=no* and *CT-truncus=no* were entered and were subsequently propagated to establish the posterior distribution for the subgoal *M1* given the new information. The Gini index of this new distribution equalled 0.37. The stopping criterion was thus not met yet for the subgoal *M1*; the same observation held with respect to the overall goal. The algorithm therefore proceeded with selecting the most informative physical test in view of the current distribution over the subgoal *M1*. The thus selected physical test was an endosonography. Its results *Endosono-loco=yes, Endosono-mediast=no, Endosono-truncus=no*, and *Endosono-wall=T3* were entered into the network and propagated. The Gini index of the distribution over *M1* now equalled 0.32. The stopping criterion was again not satisfied for the subgoal under study; the same held for the overall goal. The next physical test selected by the algorithm was an X-ray of the patient's thorax. Its result *X-lungs=yes* was entered into the network and propagated. The Gini index of the distribution over the subgoal now equalled 0.45. Note that the uncertainty with regard to the most likely value of the current subgoal *M1* again increased as a consequence of the result of the X-ray. Next, the results of the CT-scan of the thorax were prompted for. Its results *CT-lungs=yes*, *CT-loco=yes*, and *CT-organs=none* were entered into the network. After propagation, the Gini index of the distribution over the subgoal equalled 0.06. The stopping criterion was now satisfied for the subgoal *M1*. The one-step look-ahead showed that no result for the remaining physical tests could serve to increase the Gini index so as to rise above the threshold value. The stopping criterion was also met for the overall goal *Therapy*. The test-selection process was thus halted. The selection of test variables is summarised in the second column of Table 7.11. We would like to note that the sequence of physical tests selected by algorithm A_2 equals the sequence generated

Table 7.11: The sequence of tests selected for the three different algorithms.

	A_1	A_2	A_3
goal: Condition	starting package	starting package	starting package
goal: M1	*CT-liver* *CT-truncus* *Endosono-truncus* *X-lungs* *CT-lungs* **STOP**	CT-abdomen Endosonography X-thorax CT-thorax **STOP**	*CT-liver*; CT-abdomen *Endosono-truncus*; Endosonography *X-lungs*; X-thorax *CT-lungs*; CT-thorax **STOP**

by algorithm A_1. However, algorithm A_2 prompts the user for twice as many test results.

Algorithm A_3 equally showed a similar behaviour. The results from the starting package of tests were entered into the network and were subsequently propagated. With regard to the subgoal *M1*, the algorithm again selected the test variable *CT-liver* to be the most informative. Since the physical test with which a value for *CT-liver* is found also yields a value for *CT-truncus*, the user was prompted for the values of both variables. Again, the results *CT-liver=no* and *CT-truncus=no* were entered and propagated. Note that the same evidence was entered while using algorithm A_2. The value of the Gini index of the distribution over the subgoal *M1* therefore again equalled 0.37. Both the stopping criterion for the subgoal and that for the overall goal were not met and the algorithm continued the test-selection process. The test variable *Endosono-truncus* was found to be the next most informative test variable. As for algorithm A_2, the results *Endosono-loco=yes, Endosono-mediast=no, Endosono-truncus=no,* and *Endosono-wall=T3* were entered into the network and propagated, since the physical test with which a value for *Endosono-truncus* is found also yields a value for the other three test variables. Again, the Gini index over the subgoal was 0.32 and the stopping criterion was not satisfied; also the overall stopping criterion was not met. The test variable *X-lungs* was found to be the next most informative test variable. Its result *X-lungs=yes* was entered into the network and propagated. As for algorithm A_2 the Gini index over the subgoal now was 0.45 and the stopping criterion was not satisfied; also the overall stopping criterion was not met. Next, the result of the test variable *CT-lungs* was prompted for. Since the physical test with which a value for *CT-lungs* is found also yields a value for *CT-loco* and *CT-organs*, the user was prompted for the values of all three variables. The results *CT-lungs=yes, CT-loco=yes,* and *CT-organs=none* were entered into the network. After propagation, the Gini index of the distribution over the subgoal equalled 0.06. The stopping criterion was now satisfied for the subgoal *M1*. The one-step look-ahead showed that no result for the remaining test variables could serve to increase the Gini index so as to rise above the threshold value. The stopping criterion for the overall goal *Therapy* was also met. The test-selection process was thus halted. The selection of test variables is summarised in the right most column of Table 7.11. We note that the sequence of physical tests selected by algorithm A_3 equals the sequence generated by algorithms A_1 and A_2. Algorithms A_2 and A_3 prompted the user for equally many test results.

We would like to note that all three algorithms resulted in rather similar sequences of tests not just for the patient reviewed above, but also for the other seven patient cases with which

we conducted our experiment. Our findings may be explained as follows. When a single test variable is expected to result in the largest decrease of diagnostic uncertainty, then it is likely that the test to which it pertains will be the most informative test. To this end we recall that the most informative test is computed by using all combinations of possible test results for the various different variables belonging to a physical test. Now if, for instance, the variable that models the result of the CT-scan of the liver is very informative, then the CT-scan of the abdomen is likely to be informative as well. We would like to note that this property does not necessarily hold in general. Examples can be readily constructed for which the most informative test variable does not belong to the most informative physical test. However, such examples are not expected to occur very often. We note that the algorithms A_2 and A_3 especially showed the same behaviour in all patient cases. Since algorithm A_2 is computationally much more demanding than algorithm A_3, we would like to recommend the use of the semi-myopic test-selection algorithm.

To conclude, we presented the sequences of tests constructed by the algorithms A_1 and A_3 for patient 2 as discussed above, to our domain experts. We first showed the test results from the starting package of tests. For algorithm A_1, we then showed the selected test variables along with their results. For algorithm A_3, we presented the physical tests that had been selected and the results for the various different test variables that pertain to these tests. We asked the experts which sequence of tests they felt most comfortable with. We also asked them whether they would prefer to order more tests after the results of the CT-scan of the patient's lungs had become available, or would order fewer tests. The experts indicated that they felt more comfortable with the sequence of test variables generated by the semi-myopic algorithm A_3. They indicated that the sequence generated by algorithm A_1 appeared unnatural, referring to for instance the CT-scan of the liver and the CT-scan of the truncus coeliacus not being performed at the same time. They further indicated that they might have stopped the test-selection process after obtaining the positive result from the X-ray of the patient's lungs. Apparently, they felt sufficiently confident with the result of the X-ray despite its low sensitivity and specificity. This finding matches our earlier observation that our domain experts are not influenced to a large extent by the tests' characteristics. No tests were found missing from the sequence. They further indicated that the sequence of tests looked somewhat odd: they indicated that they would prefer to perform 'simple' tests first. We will address this issue in our conclusions.

7.4 Conclusions

Most test-selection algorithms currently in use select variables myopically, that is, variables are selected sequentially, on a one-by-one basis. In each step, the most informative test variable is selected, using an information measure. The user then is prompted for its results. After the results have been entered, the next test variable is selected. In non-myopic test selection, in contrast, the information gain is computed for all possible combinations of test variables. We have argued that a fully non-myopic algorithm is computationally very much demanding. Yet, we felt that myopic test selection is not very realistic for many medical applications. Our conclusion was based upon the insights that we gained from the two domain experts in the field of oesophageal cancer, with regard to their daily test-selection routines. We found for example that the test-selection process was governed by a sequence of subgoals. We further found that various test-selection variables

can pertain to just one physical diagnostic test. We also found that physical tests are ordered in packages to save the overall time of the diagnostic phase in a patient's management.

In this chapter, we focused on the first two issues described above and presented three new test-selection algorithms. Algorithm A_1 is fully myopic in nature, selecting the diagnostic tests to be performed based upon the informativity of separate variables. The difference with the basic myopic algorithm is that algorithm A_1 takes a sequence of subgoals for the selection of tests into consideration. After selecting the most informative test variable for the current subgoal, the user is prompted only for the result of this variable. Algorithm A_3 also performs the selection of tests at the level of the test variables. After selecting the most informative test variable, however, the user is prompted for the results of all test variables that belong to the same physical test as the most informative test variable. In contrast with algorithms A_1 and A_3, algorithm A_2 selects diagnostic tests at the level of physical tests; for this purpose it uses a prespecified clustering of test variables. By selecting clusters of test variables, the algorithm in essence is non-myopic in nature, yet in a very restricted sense. The new test-selection algorithms include a stopping criterion to prevent overtesting. Since the algorithms aim at test selection with regard to a sequence of subgoals in addition to an overall goal, we designed our criterion to apply to different goals. From experiments that we ran with the extended oesophageal cancer network, we concluded that a stopping criterion that builds upon the value of the Gini index of the probability distribution over a subgoal should be preferred. To prevent stopping too early, we added a limited look-ahead to our criterion.

From our experiments with the oesophageal cancer network, we found that the three algorithms behave comparably. Algorithms A_2 and A_3 in fact resulted in the same sequences of selected tests. Of these two algorithms, algorithm A_3 is to be preferred, since it is computationally much less demanding. We presented the sequences of test variables, and the moments at which the test-selection process halted, as generated by the algorithms A_1 and A_3, to our domain experts. The experts indicated that they felt more comfortable with the results of the semi-myopic algorithm A_3. They indicated, however, that in their daily routines they would prefer to perform 'simple' tests first. We would like to note that the argument of preferring simple tests, that is, tests for which the results are obtained within a short period of time or tests that have low costs, can be readily employed by our algorithm by building upon the concept of utility.

In this chapter, we focused on taking subgoals for test selection into consideration and on selecting physical tests. We argued in Chapter 6 that in our application domain tests are ordered in packages. So far, we have not addressed selecting packages of tests. Note that, in essence, we could have clustered the test variables according to the package of tests they belong to. This would have resulted in test-selection sequences that closely fit in with the daily test-selection routines of our experts. Clustering test variables to packages rather than physical tests would again increase the computation time. We would further like to note that the current test-selection routines employed by our experts are based upon general guidelines for establishing the stage of a patient's tumour and deciding upon the most preferred therapy. These guidelines, however, are designed as guidelines for all patients. Note that guidelines for specific (groups of) patients would easily become impractical to work with. With our test-selection algorithms, however, the experts are supported in working far more patient specific, which is likely to induce a decrease in the amount of money spent and an increase in the quality of their management.

CHAPTER 8

Conclusion

Decision-support systems are being developed for a wide range of domains, among which is the medical domain. We have argued that to provide true support to its users, a decision-support system should prove to be of added value, for example by resulting in lower costs, shorter waitinglists, less discomfort for patients and better care. The decision-support system should not only provide information about the most probable diseases or best suitable therapy, it should also provide its user with information about which diagnostic tests should be performed in a patient's overall management. These diagnostic tests then are performed to reduce the uncertainty about the patient's true condition. The aim of this thesis was to develop a test-selection facility for decision-support systems that include a probabilistic network for knowledge representation and reasoning. This facility should induce a test-selection strategy that is based upon mathematical principles yet fits in with the daily test-selection strategies employed by physicians. For evaluation purposes, we could build upon the oesophageal cancer network that had been developed at Utrecht University over a period of more than ten years.

We addressed the issue of modelling reliability characteristics of diagnostic tests in probabilistic networks. We argued, in Chapter 4, that a probabilistic network has to explicitly distinguish between variables that represent test results and variables that represent the underlying truth, to provide for including these characteristics. We detailed the standard characteristics, moreover, to accommodate for non-binary disease and test variables. We further stratified the characteristics for different groups of patients. We showed that the standard concepts of sensitivity and specificity can be readily reconstructed from these more detailed, and possibly stratified, characteristics.

Test selection in essence amounts to studying the informative value of the various possible test results that may be obtained. For this purpose, an information measure can be used, of which the Shannon entropy, the Gini index and the misclassification error are the most commonly employed. In Chapter 5, we analysed the differences between the three measures from a funda-

mental perspective and investigated their first derivatives. We argued that the Gini index can be looked upon as an approximation of the Shannon entropy and that the misclassification error in turn is a two-point approximation of the Gini index. With respect to their test-selection behaviour, we observed that, although a shift in many probability distributions over the diagnostic variable is valued similarly by the Gini index and the Shannon entropy, a shift to rather extreme distributions is valued much higher by the Shannon entropy than by the Gini index. Based upon this observation, the two measures are expected, at least occasionally, to select different tests. We conducted an experiment studying the behaviour of the Shannon entropy and the Gini index within the real-life setting of the oesophageal cancer network. The results from our experiment served to corroborate the differences in behaviour expected from our more fundamental analysis. A sensitivity analysis with respect to the selection of tests by the two information measures, moreover, showed that test selection based on the Shannon entropy and the Gini index is quite robust. We feel that, despite their possible differences in behaviour, both measures are equally suited for use in a decision-support system. The characteristics of the application domain under study then determine which of the two measures had best be used. We furthermore concluded from our analyses that the misclassification error should not be used for test-selection purposes.

Upon working with the oesophageal cancer network, we felt that using the sequential test-selection strategies commonly proposed in the decision-making literature would be an oversimplification of our experts' daily problem-solving practice. We acquired knowledge about the actual strategy used by the experts to provide for the design of a tailored test-selection strategy. In Chapter 6, we proposed an elicitation method composed of three focused interviews: an unstructured interview, followed up by a structured one and an interview for refinement purposes. We were able to elicit in detail the general strategy and the different arguments underlying the experts' test-selection decisions. Since we wished to ascertain that we had elicited sufficient detail for the design of an automated test selection facility, we compared the medical records of 156 patients against the elicited strategy, using the flowchart that we had constructed from the interviews. We found that the results of our comparison were influenced to a large extent by various issues. The data in a patient's record often originated from different hospitals. The data being available from different sources tended to influence the experts' selection of tests and thereby obscured their compliance with the flowchart. Moreover, our data was collected over a longer period of time during which the medical practice had changed. These changes were hard to detect from the data itself without consulting the experts. Our analysis was further seriously hampered by a relatively large number of missing values, especially since a missing value could have different meanings. As a consequence of these issues, we were not able to derive any definite conclusions from our comparison. We feel that the above and other related issues are not reserved for our field of application. In fact, we feel that these issues are likely to arise in many retrospective real-life medical data collections and should be carefully analysed.

Based upon the above insights, we designed three new algorithms for test selection and a number of associated stopping criteria. These criteria should prevent the algorithms from overtesting and prevent them from stopping the test-selection process too early. We studied the behaviour of the various stopping criteria in the context of the oesophageal cancer network and found that the simplest criteria performed best. These criteria study the probability of the most likely value of the diagnostic variable and the Gini index of the entire distribution over this vari-

able, respectively. The three test-selection algorithms have all been enhanced with a sequence of subgoals to be addressed and differ mainly in their degree of non-myopia. Our first algorithm is a fully myopic test-selection algorithm. It selects the most informative variable describing a single result from a physical test, prompts the user for the result, and continues the selection. Our fully non-myopic test-selection algorithm selects the most informative physical test, prompts the user for all its results, and continues the selection process. The third, semi-myopic test-selection algorithm selects, like the myopic algorithm, the most informative variable that describes a single result from a physical test, but prompts the user for all results from the test. We presented the test-selection sequences generated by our myopic and semi-myopic algorithms to the experts. They indicated that they felt quite comfortable with the sequence generated by especially the semi-myopic algorithm.

The goal of this thesis was to design a test-selection facility for decision-support systems with a probabilistic network. The facility should have a firm mathematical basis in decision theory yet closely fit in with daily problem-solving practice. Based upon our fundamental and experimental results, we observe that our semi-myopic test-selection algorithm which uses the Gini index as an information measure and a stopping criterion that is based upon the value of the Gini index with a one-step look-ahead facility, matches the aim that we set out to fulfil. We are aware that further research on including for instance the time it takes to obtain a test result, the costs of tests, the discomfort for the patient and the utility to perform ‘simple’ tests prior to other tests, should still be performed. We feel, however, that we have taken an important step towards the acceptance of decision-support systems as valuable assistants for physicians.

Bibliography

[1] S. Andreassen. Planning of therapy and tests in causal probabilistic networks. *Artifical Intelligence in Medicine*, 4:227 – 241, 1992.

[2] S. Andreassen, M. Suojanen, B. Falck, and K.G. Olesen. Improving the diagnostic performance of MUNIN by remodelling of the diseases. In S. Quaglini, P. Barahona, and S. Andreassen, editors, *Artificial Intelligence in Medicine. LNAI 2101*, pages 167 – 176. Springer-Verlag, 2001.

[3] S. Andreassen, M. Woldbye, B. Falck, and S.K. Andersen. MUNIN. A causal probabilistic network for interpretation of electromyographic findings. In J. McDermott, editor, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 366 – 372. Morgan Kaufmann Publishers, Los Altos, California, 1987.

[4] M. Ben-Bassat. Myopic policies in sequential classification. *IEEE Transactions on Computers*, 27(2):170 – 174, 1978.

[5] L. Breiman. Technical note: some properties of splitting criteria. *Machine Learning*, 24:41 – 47, 1996.

[6] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadworth & Brooks, Pacific Grove, 1984.

[7] R.T. Clemen. *Making Hard Decisions: An Introduction to Decision Analysis*. Duxbury Press, 1995.

[8] Commissie aanvullende diagnostistie. *Diagnostisch Kompas (in Dutch)*. College voor Zorgverzekeringen, Amstelveen, 1999.

[9] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393 – 405, 1990.

[10] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, Berlin, 1999.

[11] B. Dawson and R.G. Trapp. *Basic & Clinical Biostatistics*. McGraw-Hill, New York, 2001.

[12] F.J. Díez, J. Mira, E. Iturralde, and Z. Zubillage. DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10(1):59 – 73, 1997.

[13] P. Doubilet. A mathematical approach to interpretation and selection of diagnostic tests. *Medical Decision Making*, 3(2):177 – 195, 1983.

[14] K.A. Ericsson and H.A. Simon. *Protocol Analysis: Verbal Reports as Data, revised edition*. MIT Press, Cambridge, 1993.

[15] J.St.B.T. Evans. The knowledge elicitation problem: a psychological perspective. *Behaviour and Information Technology*, 7(2):111 – 130, 1988.

[16] W.M. Feinberg, G.W. Albers, H.J.M. Barnett, J Biller, L.R. Caplan, L.P. Carter, , R.G. Hart, R.W. Hobson, R.A. Kronmal, W.S. Moore, J.T. Robertson, H.P Adams jr, and M. Mayberg. Guidelines for the management of transient ischemic attacks. *Stroke*, 25(9):1901 – 1911, 1994.

[17] R.H. Fletcher, S.W. Fletcher, and E.H. Wagner. *Clinical Epidemiology. The Essentials*. Williams & Wilkins, Baltimore, 1996.

[18] M-F. Saint-Marc Girardin, M. Le Minor, A. Alperovitch, F. Roudot-Thoraval, J-M. Metreau, and D. Dhumeaux. Computer-aided selection of diagnostic tests in jaundiced patients. *Gut*, 26:961 – 967, 1985.

[19] P. Glasziou and J. Hilden. Test selection measures. *Medical Decision Making*, 9:133 – 141, 1989.

[20] I.J. Good and W.I. Card. The diagnostic process with special reference to errors. *Methods of Information in Medicine*, 10(3):176 – 188, 1971.

[21] G.A. Gorry, J.P. Kassirer, A. Essig, and W.B. Schwarts. Decision analysis as the basis for computer-aided management of acute renal failure. *American Journal of Medicine*, 55:473 – 484, 1973.

[22] R.A. Greenes, K.C. Cain, and C.B. Beggs. Patient-oriented performance measures of diagnostic tests. *Medical Decision Making*, 4(1):7 – 15, 1984.

[23] P.F. Griner, R.J. Mayewski, A.I. Mushlin, and P. Greenland. Selection and interpretation of diagnostic tests and procedures. *Annals of Internal Medicine*, 94(4, part 2):553 – 600, 1981.

[24] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer Verlag, New York, 2001.

[25] D. Heckerman, E. Horvitz, and B. Middleton. An approximate nonmyopic computation for value of information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):292 – 298, 1993.

[26] D.E. Heckerman and B.N. Nathwani. Toward normative expert systems. Part I: The Pathfinder project. *Methods of Information in Medicine*, 31:90 – 105, 1992.

[27] P. Jackson. *Introduction to Expert Systems*. Addison-Wesley, 1999.

[28] F.V. Jensen. *Bayesian Networks and Decision Graphs*. Statistics for Engineering and Information Science. Springer, New York, 2001.

[29] F.V. Jensen and J. Liang. drHugin - a system for value of information in bayesian networks. In *Proceedings of the Fifth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 178 – 183, 1994.

[30] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In W. Cohen and H. Hirsh, editors, *Machine Learning: Proceeding of the 11th International Conference*, pages 121–129. Morgan Kaufman Publishers, San Francisco, CA, 1994.

[31] J.P. Kassirer and R.I. Kopelman. *Learning Clinical Reasoning*. Williams & Wilkins, Baltimore, 1991.

[32] R.L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, 1993.

[33] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[34] C. Lacave and F.J. Díez. Knowledge acquisition in Prostanet, a Bayesian network for diagnosing prostate cancer. *Knowledge-Based Intelligent Information and Engineering Systems*, 2:1345 – 1350, 2003.

[35] P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 399 – 406, 1994.

[36] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50:157 – 224, 1988.

[37] P.J.F. Lucas, H. Boot, and B.G. Taal. Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, 37:206 – 219, 1998.

[38] D. Madigan and R.G. Almond. On test selection strategies for belief networks. Technical Report 20, StatSci Division, MathSoft, 1993.

[39] D. McSherry. Avoiding premature closure in sequential diagnosis. *Artificial Intelligence in Medicine*, 10:269 – 283, 1997.

[40] B. Middleton, M. Shwe, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base II. evaluation of diagnostic performance. *Methods of Information in Medicine*, 30(4):256 – 267, 1991.

[41] M.G. Morgan and M. Henrion. *Uncertainty, a Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, 1990.

[42] A.J. Neal and P.J. Hoskin. *Clinical Oncology, Basic Principles and Practice*. Arnold, Hodder Headline Group, London, 1997.

[43] A. Onisko, L. Bobrowski, M.J. Druzdzel, and H. Wasyluk. Hepar and hepar II - computer systems supporting a diagnosis of liver disorders. In *Proceedings of the Twelfth Conference on Biocybernetics and Biomedical Engineering*, pages 625 – 629, 2001.

[44] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Palo Alto, 1988.

[45] H.E. Pople. Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnosis. *Artificial Intelligence in Medicine*, pages 119 – 190, 1982.

[46] S. Quaglini, L. Dazzi, L. Gatti, M. Stefanelli, C. Fassino, and C. Tondini. Supporting tools for guideline developments and dissemination. *Artificial Intelligence in Medicine*, 14:119 – 137, 1998.

[47] S. Renooij and L.C. van der Gaag. Evidence-invariant sensitivity bounds. In M. Chickering and J. Halpern, editors, *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pages 479 – 486. AUAI Press, 1999.

[48] J.M. Schraages, S.F. Chipman, and V.L. Shelin, editors. *Cognitive Task Analysis*. Lawrence Erlbaum Associates, NJ, 2000.

[49] A.Th. Schreiber, J.M. Akkermans, A.A. Anjewierden, R. de Hoog, N.R. Shadbolt, W. Van de Velde, and B.J. Wielinga. *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press, Massachusetts, 2000.

[50] D. Sent and L.C. van der Gaag. Detailing test characteristics for probabilistic networks. In M. Dojat, E. Keravnou, and P. Barahona, editors, *Proceedings of the 9th Conference on Artificial Intelligence in Medicine in Europe, Lecture Notes in Artificial Intelligence 2780*, pages 254 – 263. Springer, 2003.

[51] D. Sent and L.C. van der Gaag. Generalised reliability characteristics for probabilistic networks. *Artificial Intelligence in Medicine*, 34:41 – 52, 2005.

[52] D. Sent and L.C. van der Gaag. On the behaviour of information measures for test selection. *submitted for publication*, 2005.

[53] D. Sent, L.C. van der Gaag, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal. On the use of vignettes for eliciting test-selection strategies. In R. Baud, M. Fieschi, P. Le Beux, and P. Ruch, editors, *The New Navigators: from Professionals to Patients; Proceedings of MIE2003*, pages 510 – 515. IOS Press, 2003.

[54] D. Sent, L.C. van der Gaag, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal. Eliciting test-selection strategies for a decision-support system in oncology. *Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, 1(6):543 – 561, 2005.

[55] B. Séroussi, J. Bouaud, and E-C Antoine. Enhancing clinical practice guidelines compliance by involving physicians in the decision process. In W. Horn, Y. Shahar, G. Lindberg, S. Andreassen, and J. Wyatt, editors, *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, pages 76 – 85. Springer, 1999.

[56] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[57] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I: the probability model and inference algorithms. *Methods of Information in Medicine*, 30:241 – 255, 1991.

[58] H.C. Sox, M.A. Blatt, M.C. Higgins, and K.I. Marton. *Medical Decision Making*. Reed Publishing, 1988.

[59] P. Szolovits. The lure of numbers: how to live with and without them in medical diagnosis. In B.E. Statland and S. Bauer, editors, *Proceedings of the Colloguium in Computer-assisted Decision Making using Clinical and Paraclinical (Laboratory) Data*, pages 65 – 75. Technico Co., 1980.

[60] L.C. van der Gaag and S. Renooij. Analysing sensitivity data. In J. Breese and D. Koller, editors, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 530 – 537. Morgan Kaufmann Publishers, San Francisco, 2001.

[61] L.C. van der Gaag, S. Renooij, B.M.P. Aleman, and B.G. Taal. Evaluation of a probabilistic model for staging of oesophageal carcinoma. In A. Hasman, B. Blobel, J. Dudeck, R. Engelbrecht, G. Gell, and H.-U. Prokosch, editors, *Medical Infobahn for Europe, Proceedings of MIE2000 and GMDS2000*, pages 772 – 776. IOS Press, Amsterdam, 2000.

[62] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal. How to elicit many probabilities. In K.B. Laskey and H. Prade, editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 647 – 654. Morgan Kaufmann Publishers, San Francisco, 1999.

[63] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal. Probabilities for a probabilistic network: A case-study in oesophageal cancer. *Artificial Intelligence in Medicine*, 25(2):123 – 148, 2002.

[64] L.C. van der Gaag, C.L.M. Witteman, , S. Renooij, and M. Egmont-Petersen. The effects of disregarding test characteristics in probabilistic networks. In S. Quaglini, P. Barahona, and S. Andreassen, editors, *Artificial Intelligence in Medicine. LNAI 2101*, pages 188 – 198. Springer-Verlag, 2001.

[65] M.W. van Someren, Y. Barnard, and J. Sandberg. *The Think Aloud Method – a Practical Approach to Modelling Cognitive Processes*. Academic Press, London, 1994.

[66] M.C. Weinstein and H.V. Fineberg. *Clinical Decision Analysis*. W.B. Saunders Company, Philadephia, 1980.

[67] W.L. Winston. *Operations Reseach: Applications and Algorithms*. Duxbury Press, 1993.

Testselectie-strategieën voor probabilistische netwerken - samenvatting

Een beslissingsondersteunend systeem is een computergebaseerd systeem dat mensen ondersteunt bij het nemen van (complexe) beslissingen. Dergelijke systemen worden vaak ontworpen om problemen op te lossen die een flinke hoeveelheid domeinkennis vereisen. In dit proefschrift richten we ons op beslissingsondersteuning binnen het domein van de geneeskunde.

De laatste jaren zijn we meer en meer gaan begrijpen van ziekteprocessen en van hoe met ziekten om te gaan. Zelfs voor gespecialiseerde artsen is het vaak moeilijk om helemaal up-to-date te blijven met de vakliteratuur, de nieuwste behandelmethoden, medicatie en ziekteleer. Aan de andere kant worden de kosten van de steeds gecompliceerder behandelingen, tests en procedures steeds hoger. Van ziekenhuizen wordt verwacht dat ze de best mogelijke zorg verlenen met een zo laag mogelijk budget. Artsen beseffen meer en meer dat als zij een fout maken, dit niet alleen gevolgen heeft voor de patiënt, maar dat de arts ook aangeklaagd kan worden met alle mogelijke financiële en professionele gevolgen vandien. Dit alles heeft ertoe geleid dat medische procedures steeds meer gestandaardiseerd worden. Beslissingsondersteunende systemen kunnen hierbij ook een belangrijke rol spelen. Zij kunnen de artsen ondersteunen met patiëntspecifieke adviezen.

Om adviezen te kunnen geven waar de arts of gebruiker vertrouwen in heeft en om daadwerkelijk een aanvulling te zijn op de kennis van de arts, moet een beslissingsondersteunend systeem aan een aantal eisen voldoen. Het systeem moet bijvoorbeeld in staat zijn om op elk moment gedurende het beslisproces informatie te geven over waarom het bepaalde conclusies heeft getrokken en waarom andere verworpen zijn. Verder zou het verschillende scenario's moeten kunnen analyseren. Verder zou het informatie moeten verschaffen over welke diagnose het meest waarschijnlijk is gegeven bepaalde testresultaten. Links met literatuur, woordenboeken en het elektronisch patiënten dossier zijn ook wenselijk. In dit proefschrift richten we ons op het uitbreiden van een beslissingsondersteunend systeem met een faciliteit die de arts kan adviseren over welke diagnostische test of combinatie van tests het best aangevraagd kan worden

voor een patiënt gegeven de informatie die op dat moment beschikbaar is. Een dergelijke faciliteit moet resulteren in een testselectie strategie die gebaseerd is op de wiskundige principes van de besliskunde, maar goed aansluit bij de testselectie strategie die de artsen hanteren in hun dagelijkse praktijk.

Om onze nieuwe faciliteit in een realistische omgeving te bestuderen en te evalueren, maken we gebruik van het slokdarmkanker netwerk. Dit beslissingsondersteunende systeem beschrijft de karakteristieken van een primaire tumor, de aan- of afwezigheid van uitzaaiingen en de eventuele doorgroei van de tumor naar naastliggende structuren. Doel van het systeem is om te bepalen wat het stadium van de kanker is. Het slokdarmkanker netwerk is een probabilistisch netwerk. Een probabilistisch netwerk bestaat uit een kwantitatief en een kwalitatief deel. Het kwantitatieve deel is een gerichte graaf waarin de knopen de belangrijke variabelen in het domein, in ons geval slokdarmkanker, modelleren en de pijlen de afhankelijkheden tussen deze variabelen aangeven. Het kwalitatieve deel is een verzameling kansverdelingen. Voor iedere variabele in de graaf zijn conditionele kansverdelingen gespecificeerd die de kans weergeven op een waarde van deze variabele gegeven alle mogelijke combinaties van waarden van de ouders van deze variabele. Probabilistische netwerken in het algemeen en het slokdarmkanker netwerk in het bijzonder, worden beschreven in hoofdstuk 2.

Een algoritme voor automatische testselectie bestaat uit drie componenten: een maat die aangeeft hoe informatief het uitvoeren van een test naar verwachting zal zijn, een zogenaamde testselectie-loop en een criterium dat aangeeft wanneer geen volgende tests meer aangevraagd zouden moeten worden, het zogenaamde stopcriterium. Testselectie algoritmen zijn myopisch of niet-myopisch. Bij myopische testselectie worden tests één voor één geselecteerd, bij niet-myopische testselectie daarentegen wordt gekeken welke test of combinatie van tests aangevraagd zou moeten worden. In hoofdstuk 3 beschrijven we verschillende maten die de informativiteit van een test weergeven. Verder bespreken we welke stopcriteria er gebruikt zijn door verschillende auteurs en welke myopische en niet-myopische testselectie algoritmen in de literatuur beschreven zijn. Ook bespreken we hier verschillende maten die informatie geven over de betrouwbaarheid van een testresultaat, zoals de sensitiviteit, die aangeeft hoe groot de kans is dat er een positief testresultaat is, gegeven de aanwezigheid van de ziekte waarop getest is, de specificiteit, de kans is dat er een negatief testresultaat is, gegeven de afwezigheid van de ziekte waarop getest is, en de predictieve waarde die informatie geeft over de kans op de aan- of afwezigheid van de ziekte waarvoor getest is gegeven een positief danwel negatief testresultaat.

De testkarakteristieken zoals hierboven beschreven zijn in de literatuur vaak gedefinieerd voor binaire variabelen die de ziekte omschrijven en binaire variabelen die testresultaten weergeven. In deze situatie is een ziekte dus simpelweg aan- of afwezig en een testresultaat positief of negatief. In een probabilistisch netwerk zitten echter vaak niet-binaire variabelen voor zowel de ziekten als de tests. Om om te kunnen gaan met niet-binaire variabelen, hebben we de standaard concepten van de sensitiviteit en de specificiteit verder gedetailleerd. We geven in hoofdstuk 4 aan hoe vanuit deze detaileerde sensitiviteiten en specificiteiten de globale sensitiviteit en specificiteit berekend kunnen worden. Soms is een testresultaat niet alleen afhankelijk van de pathologische toestand in het lichaam, maar kan deze ook afhangen van bijvoorbeeld tot welke groep patiënten iemand behoort. De sensitiviteit en specificiteit zijn nu in feite gestratificeerd voor verschillende groepen patiënten. Vanuit deze gestratificeerde testkarakteristieken kunnen

we wederom de globale sensitiviteit en specificiteit bepalen. Wanneer een testresultaat bekend is, zijn we niet langer geïnteresseerd in de sensitiviteit en de specificiteit, maar in het effect van het testresultaat op de kansverdeling over de variabele die de ziekte weergeeft. In hoofdstuk 4 laten we ook zien hoe deze zogenaamde predictieve waarde berekend kan worden gebruikmakend van de gedetailleerde sensitiviteiten en specificiteiten.

Zoals eerder aangegeven gebruikt een testselectie algoritme een maat die aangeeft hoe informatief een test is. De drie meest bekende maten zijn de Shannon entropie, de Gini index en de misclassification error. In hoofdstuk 5 analyseren we deze drie maten vanuit zowel een fundamenteel als ook een experimenteel perspectief. Allereerst hebben we gekeken naar de drie functies en hun eerste afgeleiden. We tonen aan dat de Gini index gezien kan worden als een benadering van de Shannon entropie en dat de misclassification error een tweepunts benadering is van de Gini index. Wanneer we kijken naar het testselectie gedrag van de drie maten, zien we dat voor veel kansverdelingen over de diagnostische variabele, een verschuiving in deze kansverdeling geïnduceerd door een test, door de Gini index en de Shannon entropie dezelfde waardering krijgen. Echter, voor extreme verdelingen geldt dat de Shannon entropie aan verschuivingen naar extreme kansen veel meer waarde hecht dan de Gini index. We tonen aan dat dit theoretisch aangetoonde verschil in testselectie gedrag ook te observeren is in ons systeem. De Shannon entropie en de Gini index selecteerden verschillende tests wanneer de situatie zoals hierboven beschreven zich voordeed. Ondanks dit verschil concluderen we dat beide maten geschikt zijn als maat voor testselectie in een beslissingsondersteunend systeem. We laten tevens zien dat de misclassification error eigenlijk niet geschikt is voor testselectie doeleinden omdat het de neiging heeft om tests willekeurig te kiezen wanneer alle mogelijke verschuivingen in de kansverdeling over de diagnostische variabele in hetzelfde interval zitten. Een sensitiviteitsanalyse van de Shannon entropie en de Gini index laat zien dat beide maten resulteren in robuuste selectie van tests. Wij hebben er uiteindelijk voor gekozen de Gini index te gebruiken in ons systeem.

Zoals we in hoofdstuk 3 beschreven, worden tests vaak sequentieel geselecteerd. Een test wordt aangevraagd, het resultaat wordt ingevoerd en een volgende test wordt aangevraagd. Omdat wij van mening waren dat een dergelijke testselectie strategie een oversimplificatie zou zijn van het dagelijkse probleemoplossen van onze domein deskundigen, hebben we besloten om kennis over hun strategie te eliciteren om zo een testselectie faciliteit te kunnen maken die beter aansluit bij de dagelijkse praktijk. In hoofdstuk 6 beschrijven we hoe we de testselectie strategie die onze experts dagelijks hanteren hebben geëliciteerd. We hebben hiervoor gebruik gemaakt van drie interviews: een ongestructureerd interview, gevolgd door een gestructureerd interview en als laatste een interview om onze resultaten te bespreken en te verfijnen. Op basis van het eerste interview, dat bestond uit vragen als “beschrijf welke tests je aanvraagt vanaf het moment dat een patiënt voor het eerst bij je komt tot aan het vaststellen van de juiste behandeling”, hebben we geconcludeerd dat tests aangevraagd worden in drie verschillende pakketjes. Alle tests uit één pakket worden tegelijkertijd aangevraagd. Op basis van deze kennis hebben we een eenvoudig stroomdiagram opgesteld. Tijdens het tweede interview hebben we acht patiëntencases besproken die door onszelf samengesteld waren. Deze acht cases bevonden zich op de grenzen van de verschillende beslismomenten in ons stroomdiagram. Deze cases waren weergegeven op kaartjes, waarbij ieder kaartje een pakket tests aangaf. Op basis van dit tweede interview hebben we ons stroomdiagram uitgebreid en aangepast. Tijdens het derde en laatste interview hebben we

het opgestelde stroomdiagram besproken aan de hand van een data-analyse die we uitgevoerd hadden en her en der kleine verfijningen aangebracht.

We wilden er zeker van zijn dat we voldoende details hadden geëliciteerd om een goede automatische testselectie faciliteit te kunnen maken. We hebben daartoe de gegevens van 156 patiënten vergeleken met het stroomdiagram dat we opgesteld hadden. We vonden echter dat onze resultaten sterk beïnvloed werden door verschillende zaken. Allereerst was de data afkomstig uit meerdere ziekenhuizen, wat de selectie van tests beïnvloedde en daardoor niet altijd overeenkwam met het opgestelde stroomdiagram. Verder was de data verzameld gedurende een lange periode. In deze tijdsperiode is de medische praktijk wat betreft het diagnostiseren van slokdarmkanker althans, aangepast. Tenslotte werd onze analyse sterk bemoeilijkt door een grote hoeveelheid missende waarden. Door dit alles waren we niet in staat om conclusies te trekken uit onze vergelijking.

Op basis van de resultaten uit de eerdere hoofdstukken, hebben we drie nieuwe testselectie algoritmen ontworpen. Deze worden beschreven in hoofdstuk 7. Tevens beschrijven we daar de verschillende stopcriteria die we ontworpen hebben. Deze criteria zouden moeten voorkomen dat teveel tests aangevraagd worden, maar, nog belangrijker, dan er niet te vroeg gestopt wordt met het aanvragen van diagnostische tests. Het gedrag van de verschillende stopcriteria is bekeken aan de hand van hun gedrag in de context van het slokdarmkanker netwerk. We concluderen in hoofdstuk 7 dat de eenvoudigste criteria het beste functioneren. Deze criteria bekijken de kans van de meest waarschijnlijke waarde van de diagnostische variabele respectievelijk de Gini index van de kansverdeling over deze variabele. Ons uiteindelijke stopcriterium is een combinatie van deze twee criteria.

De drie testselectie algoritmen maken gebruik van de subdoelen, weergegeven als beslismomenten in het eerder beschreven stroomdiagram. Het eerste algoritme selecteert een test variabele op basis van de Gini index, vult de waarde van deze variabele in, en bepaalt vervolgens welke test variabele nu het meest informatief is. Het tweede algoritme daarentegen selecteert de meest informatieve werkelijke test en vult de waarden van alle test variabelen die bij deze fysieke test horen in. Het derde, semi-myopische testselectie algoritme selecteert de meest informatieve test variabele zoals ook ons eerste algoritme deed, maar vult niet alleen de waarde van deze test variabele in, maar ook die van alle variabelen die tot dezelfde fysieke test behoren als de test variabele die geselecteerd was. De testselectie strategie die door deze drie algoritmen gevonden werd is voorgelegd aan de experts. Deze waren van mening dat ze zich met name prettig voelden bij de resultaten van het derde algoritme.

In Hoofstuk 8 geven we tenslotte aan dat het doel van dit proefschrift was om een testselectie faciliteit te ontwikkelen voor beslissingsondersteunende systemen met een probabilistisch netwerk. De basis van deze faciliteit moest in de wiskunde liggen, echter de faciliteit zou moeten resulteren in testselectie strategieën die aansluiten bij de dagelijkse praktijk. Gebaseerd op onze fundamentele en experimentele resultaten, kunnen we concluderen dat ons semi-myopische testselectie algoritme dat de Gini index gebruikt als informatiemaat en een stopcriterium dat gebaseerd is op waarde van de Gini index met een zogenaamde one-step look-ahead faciliteit, aansluit bij het doel dat we gesteld hadden. Verder onderzoek zou zich moeten richten op het meenemen van de tijd die het kost om een resultaat te verkrijgen van een test, de kosten van een test, patiëntenvoorkeuren en de utiliteit om simpele tests uit te voeren voordat je overgaat

naar de meer gecompliceerde tests. Ondanks dat deze zaken duidelijk nog aandacht verdienen, zijn wij van mening dat we reeds nu een belangrijke stap hebben gezet in de acceptatie van beslissingsondersteunende systemen als waardevolle assistenten van artsen.

Dankwoord

Eindelijk! Het is af! De afgelopen jaren hebben veel mensen bewust of onbewust een steentje bijgedragen aan de totstandkoming van dit proefschrift. Ik vind het een eer dat ik jullie mag vermelden in mijn proefschrift. Helaas kan ik niet iedereen bedanken. Nadia, ook al ben je dan niet meer bij ons, ik weet zeker dat je me geholpen hebt en dat ook jij nu trots neer zult kijken!

Mijn dank gaat natuurlijk uit naar mijn promotor Linda van der Gaag. Al tijdens mijn studie viel je op door je enorme gedrevenheid, je kennis en aanstekelijke enthousiasme. Vanaf het moment dat ik bij jou als AiO begon, heb je me gesteund en geholpen bij iedere stap die ik zette. Je inzet bij ons onderzoek en bij iedere letter die op papier zette is meer dan bewonderswaardig. Je leerde me preciezer te zijn, schaafde mijn kennis bij en steunde me op tijden dat ik het even wat minder zag zitten. Waar ik je ook dankbaar voor ben zijn de leuke uitstapjes en gesprekken die we gehad hebben. Vooral het uitstapje met rondvliegende salami zal me nog lang bijstaan. Dank je voor alles!

Jan van Leeuwen wil ik graag bedanken, omdat hij mogelijk gemaakt heeft dat ik mijn onderzoek nog een paar maanden langer mocht uitvoeren. Dank voor het begrip! Verder natuurlijk ook Berthe Aleman en Babs Taal van het Nederlands Kanker Instituut/Antoni van Leeuwenhoekhuis. Dank jullie voor jullie tijd en geduld tijdens de interviews die we met jullie hadden en het delen van jullie kennis met ons. Door jullie hulp hebben we dit resultaat kunnen neerzetten. Verder natuurlijk ook Cilia Witteman voor jouw bijdrage aan de interviews en het resulterende hoofdstuk in mijn proefschrift. Programmeren heb ik nooit leuk gevonden en dankzij de inzet van drie fantastische kanjers heb ik dat gelukkig tijdens dit onderzoek kunnen vermijden. Arend-Jan de Groote, Arjan van IJzendoorn en Martijn Schrage wil ik danken voor het implementeren van de vele experimenten die we bedachten. Arjan en Martijn: bedankt voor het ontwikkelen van Dazzle. Dat maakte het leven een stuk aangenamer. Arjan, jou heb ik meer dan eens lastig gevallen omdat ik het toch nèt weer iets anders wilde dan je gedaan had, of omdat ik toch nog nèt iets wilde uitbreiden. Zonder mokken ging je altijd maar weer aan de slag en dacht je met me mee. De snelheid waarmee je met resultaten kwam, was verbluffend. Ik hoop dat jullie nog lang met veel plezier aan Dazzle zullen werken! Arend, je verliet onze groep en de kamer die we deelden

omdat je je ging richten op je studie geneeskunde. Ik ben ervan overtuigd dat je een goed arts zult worden! In ieder geval eentje met een gezonde dosis humor. Jouw gezelschap was altijd aangenaam. Vooral onze partijtjes backgammon tegen elkaar waren meer dan leuk! Dat maakte de dag weer helemaal compleet. Zonder jou was het toch wennen. Peter Bosman wil ik graag bedanken voor zijn hulp met het plotten van plaatjes. Silja Renooij verdient ook een meer dan eervolle vermelding. Dank je voor alle hulp bij het lay-outen van mijn proefschrift, bij het maken van plaatjes, en ga zo maar door. Waar ik je in het bijzonder voor wil danken is dat je de concept versie van mijn proefschrift zo grondig hebt doorgelezen en alle dingetjes die je tegenkwam zo zorgvuldig aan me doorgaf. Ad Feelders en Eveline Helsper wil ik graag danken voor hun commentaar bij respectievelijk hoofdstuk 5 en 6. Petra Geenen wil ik graag bedanken voor het organiseren van een proefverdediging en de overige mensen in de DSS-groep voor het lezen van het proefschrift en het bedenken van vragen. Jullie waren fijne collega's! For reviewing this thesis, I would like to thank Stig Andersen, Yolanda van der Graaf, John-Jules Meyer, Silvana Quaglini and Arno Siebes. Fijne collega's zijn heel belangrijk, dat weet iedereen. Zeker als het wat minder gaat, is het fijn als je op mensen kunt leunen. Ik wil daarom speciaal Monique, Floor en Lydia bedanken. Dank jullie voor jullie luisterend oor, de kopjes thee, de snoepjes bij de balie toen ik nog het genoegen had om daar tegenover te zitten en jullie constante interesse in jullie medemens! Ik zal jullie missen!

Onderzoek doen en een proefschrift schrijven gaat met pieken en dalen. Daarom wil ik ook graag mensen bedanken die het leven buiten de universiteit aangenamer maakten. Allereerst natuurlijk mijn paranimfen Annemarie en Margot. Anne, we kennen elkaar al weer heel wat jaren. Nooit ben ik vergeten hoe ik, amper een paar weken op mijn nieuwe school, de ziekte van pfeiffer kreeg. Je kende me nog maar net en toch kwam je vaak langs en mocht ik je aantekeningen hebben. Gelukkig zijn we elkaar daarna niet uit het oog verloren. Ik vind het fijn dat jij mij straks bij de promotie gaat bijstaan zoals je dat al jaren gedaan hebt. Onze vriendschap is mij heel dierbaar. Margot, we kennen elkaar van het bestuur van de studentenruiters. Ook alweer een hele tijd dus. We hebben al zoveel meegemaakt samen. Mede door jouw steun ben ik er weer bovenop gekomen. Het wordt tijd dat we een periode gaan afsluiten en vol goede moed een nieuwe periode in ons leven gaan inslaan. Fijn dat ook jij erbij wilt zijn.

Mijn overige bestuursgenoten Anita, Caroline, Erik, Joost, Lia, Monique en Wouter wil ik ook bedanken. De tijd bij de studentenruiters waarbij ik mijzelf en anderen er keer op keer van probeerde te overtuigen dat je als AiO ook best nog voor student kunt doorgaan, zal ik niet snel vergeten. Onze vergaderingen waren, zoals Joost het menigmaal verwoordde erg gezellig. Joost, met name jou wil ik bedanken voor de vele goede gesprekken onder het genot van een of meerdere biertjes. Jouw geloof in mij deed mij besluiten door te gaan, dingen aan te pakken en soms wegen te bewandelen die helemaal nieuw voor me waren. Nu is dan toch de tijd aangebroken dat ik de studentenruiters vaarwel ga zeggen. Al hoop ik niet dat dat betekent dat we elkaar uit het oog verliezen! Thijs, wij leerden elkaar kennen in de tijd bij Hippeia. Ik zou er een boek over kunnen schrijven! Dank je voor de vele uren gezelligheid en steun. Ik ben trots dat ik jou mijn "broer" mag noemen...

Jeroen, wij leerden elkaar onder minder prettige omstandigheden kennen. Jouw altijd relativerende woorden, de vele telefoongesprekken die we hadden en jouw enthousiasme als ik vertelde dat ik weer een stapje dichterbij het voltooien van het proefschrift was, waren fantastisch. Die

aanbieding van het zeilen staat nog? Natuurlijk ook een plekje voor de Cambridgelaan en de Cambridgebar in het bijzonder. Claudia, bedankt voor het eindeloos voorlezen van getalletjes als ik weer eens tabellen zat ik te typen of dingen zat te turven! Désirée, onze kaartspelletjes tot diep in de nacht en het nasynchroniseren van foute reclames waren fantastisch. Net wat ik op dat moment nodig had! Anoek, jou wil ik bedanken voor het mede-ontwerpen van de kaft! Ik sta er elke keer weer verbaasd van hoe jij ideeën van mensen zo duidelijk op papier weet te krijgen. We hebben eindeloos over de inhoud van het proefschrift gepraat en wat mijn idee bij de kaft was. Het resultaat mag er wezen! Fleur, jij bedankt voor de vele overheerlijke maaltijden die je mij voorschotelde de afgelopen jaren. Vergeet niet die lekkere recepten nog een keer te geven hè! Verder natuurlijk alle andere mensen in “de bar” waar ik lief en leed mee gedeeld heb.

Ik wil mijn ouders, mam en Ben, pa en Erna bedanken voor hun vertrouwen in mij en in mijn kunnen. Jullie hebben mij altijd alle ruimte en mogelijkheden geboden om daar te komen waar ik nu ben. Dank! Esther-Mirjam, ik ben blij dat je weer in Nederland bent! Ik ben je dankbaar voor al je goede raad de laatste tijd! Ook mijn nichtje Luna verdient een plaatsje. Jij bent nu al heel bijzonder! Lieve Geert, hoe zeer ik ook probeer de juiste woorden te vinden, het zal nooit beschrijven wat jij voor mij betekent. We leerden elkaar bijna drie jaar geleden kennen en al snel hadden we urenlange serieuze gesprekken en daarnaast ook veel lol. Geen van beiden moesten we eraan denken dat we het ooit zonder elkaars gezelschap zouden moeten stellen. Nu zijn we inmiddels allebei weer een fase verder. Jij bent inmiddels afgestudeerd en ook ik ga een nieuwe fase in. Ik wil je bedanken voor je bijdrage aan mijn proefschrift. Je las alles, hielp mee met turven, je hielp me met de medische stukken, kookte en ga zo maar door. Maar bovenal dank ik je omdat jij jij bent en voor alles wat je voor me doet.

Curriculum Vitae

Daniëlle Sent

21 november 1977

Geboren te Rheden.

september 1990 – juni 1996

Gymnasium aan de Katholieke Scholengemeenschap De Grundel in Hengelo (O); Eerste vier jaar aan de Regionale Scholengemeenschap Het Rhedens in Rozendaal.
Diploma behaald in juni 1996.

september 1996 – april 2000

Studie Informatica aan het Informatica Instituut van de Universiteit Utrecht.
Doctoraal diploma behaald in april 2000.

mei 2000 – april 2005

Assistent in Opleiding bij het Instituut voor Informatica en Informatiekunde van de Universiteit Utrecht.

Titles in the SIKS Dissertation Series

98–1 Johan van den Akker, *DEGAS – An Active, Temporal Database of Autonomous Objects.*

98–2 Floris Wiesman, *Information Retrieval by Graphically Browsing Meta-Information.*

98–3 Ans Steuten, *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective.*

98–4 Dennis Breuker, *Memory versus Search in Games.*

98–5 Eduard Oskamp, *Computerondersteuning bij Straftoemeting.*

99–1 Mark Sloof, *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products.*

99–2 Rob Potharst, *Classification using decision trees and neural nets.*

99–3 Don Beal, *The Nature of Minimax Search.*

99–4 Jacques Penders, *The practical Art of Moving Physical Objects.*

99–5 Aldo de Moor, *Empowering Communities: A Method for the Legitimate User–Driven Specification of Network Information Systems.*

99–6 Niek Wijngaards, *Re-design of compositional systems.*

99–7 David Spelt, *Verification support for object database design.*

99–8 Jacques Lenting, *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.*

2000–1 Frank Niessink, *Perspectives on Improving Software Maintenance.*

2000–2 Koen Holtman, *Prototyping of CMS Storage Management.*

2000–3 Carolien Metselaar, *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief.*

2000–4 Geert de Haan, *ETAG, A Formal Model of Competence Knowledge for User Interface Design.*

2000–5 Ruud van der Pol, *Knowledge-based Query Formulation in Information Retrieval.*

2000–6 Rogier van Eijk, *Programming Languages for Agent Communication.*

2000–7 Niels Peek, *Decision-theoretic Planning of Clinical Patient Management.*

2000–8 Veerle Coupé, *Sensitivity Analyis of Decision-Theoretic Networks.*

2000–9 Florian Waas, *Principles of Probabilistic Query Optimization.*

2000–10 Niels Nes, *Image Database Management System Design Considerations, Algorithms and Architecture.*

2000–11 Jonas Karlsson, *Scalable Distributed Data Structures for Database Management.*

2001–1 Silja Renooij, *Qualitative Approaches to Quantifying Probabilistic Networks.*

2001–2 Koen Hindriks, *Agent Programming Languages: Programming with Mental Models.*

2001–3 Maarten van Someren, *Learning as Problem Solving.*

2001–4 Evgueni Smirnov, *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets.*

2001–5 Jacco van Ossenbruggen, *Processing Structured Hypermedia: A Matter of Style.*

2001–6 Martijn van Welie, *Task-based User Interface Design.*

2001–7 Bastiaan Schonhage, *Diva: Architectural Perspectives on Information Visualization.*

2001–8 Pascal van Eck, *A Compositional Semantic Structure for Multi-Agent Systems Dynamics.*

2001–9 Pieter Jan 't Hoen, *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes.*

2001–10 Maarten Sierhuis, *Modeling and Simulating Work Practice BRAHMS: a Multiagent Modeling and Simulation Language for Work Practice Analysis and Design.*

2001–11 Tom M. van Engers, *Knowledge Management: The Role of Mental Models in Business Systems Design.*

2002–1 Nico Lassing, *Architecture-Level Modifiability Analysis.*

2002–2 Roelof van Zwol, *Modelling and Searching Web-based Document Collections.*

2002–3 Henk Ernst Blok, *Database Optimization Aspects for Information Retrieval.*

2002–4 Juan Roberto Castelo Valdueza, *The Discrete Acyclic Digraph Markov Model in Data Mining.*

2002–5 Radu Serban, *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents.*

2002–6 Laurens Mommers, *Applied legal epistemology; Building a Knowledge-based Ontology of the Legal Domain.*

2002–7 Peter Boncz, *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications.*

2002–8 Jaap Gordijn, *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas* .

2002–9 Willem-Jan van den Heuvel, *Integrating Modern Business Applications with Objectified Legacy Systems.*

2002–10 Brian Sheppard, *Towards Perfect Play of Scrabble.*

2002–11 Wouter C.A. Wijngaards, *Agent Based Modelling of Dynamics: Biological and Organisational Applications.*

2002–12 Albrecht Schmidt, *Processing XML in Database Systems.*

2002–13 Hongjing Wu, *A Reference Architecture for Adaptive Hypermedia Applications.*

2002–14 Wieke de Vries, *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems.*

2002–15 Rik Eshuis, *Semantics and Verification of UML Activity Diagrams for Workflow Modelling.*

2002–16 Pieter van Langen, *The Anatomy of Design: Foundations, Models and Applications.*

2002–17 Stefan Manegold, *Understanding, Modeling, and Improving Main-Memory Database Performance.*

2003–1 Heiner Stuckenschmidt, *Ontology-Based Information Sharing in Weakly Structured Environments.*

2003–2 Jan Broersen, *Modal Action Logics for Reasoning About Reactive Systems.*

2003–3 Martijn Schuemie, *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy.*

2003–4 Milan Petkovic, *Content-Based Video Retrieval Supported by Database Technology.*

2003–5 Jos Lehmann, *Causation in Artificial Intelligence and Law - A modelling approach.*

2003–6 Boris van Schooten, *Development and Specification of Virtual Environments.*

2003–7 Machiel Jansen, *Formal Explorations of Knowledge Intensive Tasks.*

2003–8 Yongping Ran, *Repair Based Scheduling.*

2003–9 Rens Kortmann, *The Resolution of Visually Guided Behaviour.*

2003–10 Andreas Lincke, *Electronic Business Negotiation: Some Experimental Studies on the Interaction between Medium, Innovation Context and Culture.*

2003–11 Simon Keizer, *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks.*

2003-12 Roeland Ordelman, *Dutch Speech Recognition in Multimedia Information Retrieval.*

2003–13 Jeroen Donkers, *Nosce Hostem - Searching with Opponent Models.*

2003–14 Stijn Hoppenbrouwers, *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations.*

2003–15 Mathijs de Weerdt, *Plan Merging in Multi-Agent Systems.*

2003–16 Menzo Windhouwer, *Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses.*

2003–17 David Jansen, *Extensions of Statecharts with Probability, Time, and Stochastic Timing.*

2003–18 Levente Kocsis, *Learning Search Decisions.*

2004–1 Virginia Dignum, *A Model for Organizational Interaction: Based on Agents, Founded in Logic.*

2004–2 Lai Xu, *Monitoring Multi-party Contracts for E-business.*

2004–3 Perry Groot, *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving.*

2004–4 Chris van Aart, *Organizational Principles for Multi-Agent Architectures.*

2004–5 Viara Popova, *Knowledge Discovery and Monotonicity.*

2004–6 Bart-Jan Hommes, *The Evaluation of Business Process Modeling Techniques.*

2004–7 Elise Boltjes, *Voorbeeldig Onderwijs; Voorbeeldgestuurd Onderwijs, een Opstap naar Abstract Denken, Vooral voor Meisjes.*

2004–8 Joop Verbeek, *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale Politiële Gegevensuitwisseling en Digitale Expertise.*

2004–9 Martin Caminada, *For the Sake of the Argument; Explorations into Argument-based Reasoning.*

2004–10 Suzanne Kabel, *Knowledge-rich Indexing of Learning-objects.*

2004–11 Michel Klein, *Change Management for Distributed Ontologies.*

2004–12 The Duy Bui, *Creating Emotions and Facial Expressions for Embodied Agents.*

2004–13 Wojciech Jamroga, *Using Multiple Models of Reality: On Agents who Know how to Play.*

2004–14 Paul Harrenstein, *Logic in Conflict. Logical Explorations in Strategic Equilibrium.*

2004–15 Arno Knobbe, *Multi-Relational Data Mining.*

2004–16 Federico Divina, *Hybrid Genetic Relational Search for Inductive Learning.*

2004–17 Mark Winands, *Informed Search in Complex Games.*

2004–18 Vania Bessa Machado, *Supporting the Construction of Qualitative Knowledge Models.*

2004–19 Thijs Westerveld, *Using Generative Probabilistic Models for Multimedia Retrieval.*

2004–20 Madelon Evers, *Learning from Design: Facilitating Multidisciplinary Design Teams.*

2005–1 Floor Verdenius, *Methodological Aspects of Designing Induction-Based Applications.*

2005–2 Erik van der Werf, *AI Techniques for the Game of Go.*

2005–3 Franc Grootjen, *A Pragmatic Approach to the Conceptualisation of Language.*

2005–4 Nirvana Meratnia, *Towards Database Support for Moving Object Data.*

2005–5 Gabriel Infante-Lopez, *Two-Level Probabilistic Grammars for Natural Language Parsing.*

2005–6 Pieter Spronck, *Adaptive Game AI.*

2005–7 Flavius Frasincar, *Hypermedia Presentation Generation for Semantic Web Information Systems.*

2005–8 Richard Vdovjak, *A Model-driven Approach for Building Distributed Ontology-based Web Applications.*

2005–9 Jeen Broekstra, *Storage, Querying and Inferencing for Semantic Web Languages.*

2005–10 Anders Bouwer, *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments.*

2005–11 Elth Ogston, *Agent Based Matchmaking and Clustering - A Decentralized Approach to Search.*

2005–12 Csaba Boer, *Distributed Simulation in Industry.*

2005–13 Fred Hamburg, *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen.*

2005–14 Borys Omelayenko, *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics.*

2005–15 Tibor Bosse, *Analysis of the Dynamics of Cognitive Processes.*

2005–16 Joris Graaumans, *Usability of XML Query Languages.*

2005–17 Boris Shishkov, *Software Specification Based on Re-usable Business Components.*

2005–18 Daniëlle Sent, *Test-selection Strategies for Probabilistic Networks.*