ELSEVIER Artificial Intelligence in Medicine 25 (2002) 123–148

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Probabilities for a probabilistic network: a case study in oesophageal cancer

L.C. van der Gaag[a,*], S. Renooij[a], C.L.M. Witteman[a], B.M.P. Aleman[b], B.G. Taal[b]

[a]*Institute of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands*

[b]*Department of Radiation Oncology and Gastroenterology, The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands*

Received 17 January 2001; received in revised form 10 September 2001; accepted 2 October 2001

Abstract

With the help of two experts in gastrointestinal oncology from The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, a decision-support system is being developed for patient-specific therapy selection for oesophageal cancer. The kernel of the system is a probabilistic network that describes the presentation characteristics of cancer of the oesophagus and the pathophysiological processes of invasion and metastasis. While the construction of the graphical structure of the network was relatively straightforward, probability elicitation with existing methods proved to be a major obstacle. To overcome this obstacle, we designed a new method for eliciting probabilities from experts that combines the ideas of transcribing probabilities as fragments of text and of using a scale with both numerical and verbal anchors for marking assessments. In this paper, we report experiences with our method in eliciting the probabilities required for the oesophagus network. The method allowed us to elicit many probabilities in reasonable time. To gain some insight in the quality of the probabilities obtained, we conducted a preliminary evaluation study of our network, using data from real patients. We found that for 85% of the patients, the network predicted the correct cancer stage.

Keywords: Elicitation of judgemental probabilities; Probabilistic networks

1. Introduction

The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, is a specialised centre for the treatment of cancer patients. Every year some 80 patients receive treatment for

[*] Corresponding author.
E-mail address: linda@cs.uu.nl (L.C. van der Gaag).

0933-3657/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0933-3657(02)00012-X

oesophageal cancer at the centre. These patients are currently assigned to a therapy by means of a standard protocol that includes a small number of prognostic factors. Based upon this protocol, 75% of the patients show a favourable response to the therapy provided. One out of every four patients, however, develops more or less serious complications as a result of the therapy. To arrive at a more fine-grained protocol with a more favourable response rate, a decision-support system is being developed for patient-specific therapy selection. The system is destined for use in clinical practice and is being constructed with the help of two experts in gastrointestinal oncology from The Netherlands Cancer Institute. The two experts are the co-authors B.M.P. Aleman and B.G. Taal of the present paper.

The kernel of our decision-support system is a probabilistic network. A probabilistic network is a mathematical model that encodes statistical variables and the probabilistic relationships between them in a graphical structure; the strengths of the relationships between the variables are indicated by conditional probabilities [7]. The probabilistic network of our system models the presentation characteristics of an oesophageal tumour, such as its length and shape, as well as the pathophysiological processes underlying its invasion into the oesophageal wall and its metastasis. The network further captures the sensitivity and specificity characteristics of the diagnostic tests that are typically performed to assess the stage of a patient's cancer. For prognostication, the network in addition describes the possible effects of the different therapies available. When a patient's symptoms and test results are entered, the network predicts the most likely stage of the patient's cancer and assesses the most likely outcomes of the different treatment alternatives. In the sequel, we will use the phrase *oesophagus network* to refer to our probabilistic network of oesophageal cancer.

The oesophagus network is being constructed with the help of two domain experts. First, we carefully modelled, in the network's graphical structure, the relationships between the statistical variables that represent for example the characteristics of an oesophageal tumour and the possible effects of the different therapies. We then focused on obtaining the probabilities required for the quantitative part of the network. This task is generally acknowledged to be the most daunting in the construction of a probabilistic network [5]. For our network, it indeed turned out to be the hardest and most time consuming of the various tasks involved. At first sight, many sources of probabilistic information appeared to be readily available. Unfortunately, a thorough literature review did not yield any usable results. Moreover, we were not able to compose a rich enough data collection from which the probabilities could reliably be estimated. The single remaining source of probabilistic information, therefore, was the knowledge and personal clinical experience of our two domain experts.

The problems that are typically encountered when eliciting probabilities from human domain experts are widely known [8]. An expert's assessments may for example reflect various biases and may not be properly calibrated. Acknowledging these problems, in the field of decision analysis several methods have been developed for eliciting judgemental probabilities, ranging from probability scales for marking assessments to gambles [10,13]. For eliciting the probabilities required for the oesophagus network, we decided to use these well-known methods with our experts. Unfortunately, we encountered numerous problems. Most importantly, we found that using the more involved methods tended to take considerable time with every single assessment. In fact, it became clear that, with these methods, the elicitation of the large number of probabilities required for our network was infeasible. We concluded that existing elicitation methods may work well for small numbers of probabilities, but do not easily scale up to the thousands of probabilities that are typically required for even a moderately sized probabilistic network. Building upon our negative experiences, we designed a new method that we tailored to the elicitation of a large number of probabilities. Our method combines several ideas, such as transcribing the required probabilities as fragments of text and providing a scale with both numerical and verbal anchors for marking assessments. We used our new method for the elicitation of the probabilities for the oesophagus network. With the method, our domain experts provided the probabilities required at a rate of over 150 numbers per hour.

To gain some insight in the quality of the probabilities that we obtained with our new elicitation method, we conducted a preliminary evaluation study of the oesophagus network, using data from real patients diagnosed with oesophageal cancer. We focused on the part of the network that provides for establishing the stage of a patient's cancer. This stage summarises the depth of invasion of the primary tumour into the oesophageal wall and the extent of its metastasis, and is indicative of the prognosis for the patient. We would like to note that in our decision-support system the depth of invasion and extent of metastasis themselves are of interest rather than the stage derived from them. Focusing on the summarising stage, however, serves to provide overall insight in the diagnostic part of the network. We found that for 85% of the patients, the stage yielded by our network as the most likely stage matched the stage that was recorded in the patient's data.

In this paper, we describe our new method for probability elicitation and report experiences with the method in eliciting from our domain experts the probabilities required for the oesophagus network. In Section 2 we describe the network. In Section 3 we discuss our initial experiences with probability elicitation using existing methods. In Section 4 we detail the method that we designed for eliciting a large number of probabilities. In Section 5 we describe our experiences with this method in the construction of the quantitative part of the oesophagus network; more specifically, we comment on the observations made by the domain experts. In Section 6 we reflect on the probabilities obtained and present the results of a preliminary evaluation study of our network. The paper ends with some concluding observations in Section 7.

2. The oesophagus network and the patient data

With the help of two experts in gastrointestinal oncology from The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, we constructed a probabilistic network for oesophageal cancer. In this section, we provide some background knowledge on cancer of the oesophagus and introduce the network. In addition, we briefly describe the patient data that we used in our preliminary evaluation study.

2.1. Oesophageal cancer

As a consequence of a lesion of the oesophageal wall, for example as a result of frequent reflux or associated with smoking and drinking habits, a tumour may develop in a patient's oesophagus. The various presentation characteristics of the tumour, which include its location in the oesophagus and its histological type, length, and macroscopic shape, influence its prospective growth. The tumour typically invades the oesophageal wall and upon further growth may invade such neighbouring structures as the trachea and bronchi or the diaphragm, dependent upon its location in the oesophagus. In time, the tumour may give rise to lymphatic metastases in distant lymph nodes and to haematogenous metastases in, for example, the lungs and the liver. The depth of invasion and extent of metastasis, summarised in the cancer's stage, largely influence a patient's life expectancy and are indicative of the effects and complications to be expected from the different available treatment alternatives. To establish these factors in a patient, typically a number of diagnostic tests are performed, ranging from multiple biopsies of the primary tumour to a gastroscopic and endosonographic examination of the oesophagus and a CT-scan of the patient's chest and liver.

While establishing the presence of an oesophageal tumour in a patient is relatively straightforward, the staging of the cancer and especially the selection of an appropriate therapy are far harder tasks. In The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, different treatment alternatives are available, ranging from surgical removal of the oesophagus to positioning a prosthesis. The effects aimed at by providing a therapy include removal or reduction of the patient's primary tumour to prolong life expectancy and to improve passage of food through the oesophagus. The therapies differ in the extent to which these effects can be attained, however. For example, where the main goal of surgical removal of the oesophagus is to attain a better life expectancy for a patient, positioning a prosthesis in the oesophagus cannot improve life expectancy: the latter is performed merely to relieve the patient's difficulty with swallowing food. Providing a therapy is often accompanied not just by beneficial effects but also by complications. These complications can be very serious and may in fact result in death. The effects and complications expected from the different therapies for a specific patient depend on the characteristics of his or her primary tumour, on the depth of invasion of the tumour into the oesophageal wall and neighbouring structures, and on the extent of metastasis of the cancer. The cancer's stage therefore plays a crucial role in the selection of an appropriate therapy for a patient.

2.2. The oesophagus network

We captured the state-of-the-art knowledge about oesophageal cancer and its treatment in a *probabilistic network*, also known as a Bayesian network or causal network [7]. The network includes a graphical structure encoding statistical variables and the probabilistic relationships between them. Each variable represents a diagnostic or prognostic factor that is relevant for establishing the stage of a patient's cancer or for predicting the outcome of treatment. The probabilistic influences among the variables are represented by directed links; the strengths of these influences are indicated by conditional probabilities. Our probabilistic network of oesophageal cancer currently includes over 70 statistical variables and more than 4000 conditional probabilities. The graphical structure and its associated probabilities uniquely capture a joint probability distribution over the represented variables. Any probability of interest over these variables can therefore be computed from the network. More specifically, the stage of a patient's cancer can be established by entering his or her symptoms and test results into the network, and computing the effect of these observations on the marginal probability distribution for the variable that models the cancer's stage.

Thus far, we focused our elicitation efforts on the part of the network that pertains to the characteristics, depth of invasion, and metastasis of an oesophageal tumour. This part constitutes a coherent and self-contained probabilistic network. In the sequel, we will refer to this network by the phrase *oesophagus network* as well, as long as ambiguity cannot occur. The network's graphical structure is depicted in Fig. 1; the figure also shows the prior marginal probability distributions for the various statistical variables.

The 40 variables involved required some 1000 probability assessments. The variable requiring the largest number of assessments, 144, models the cancer's stage. This variable classifies a patient's oesophageal cancer in one of six categories of disease. It is deterministic in the sense that its value is determined uniquely by the values of its predecessors in the graphical structure of the network; the probabilities required for this variable therefore are all equal to 0 or 1. The non-deterministic variable requiring the largest number of probability assessments is the variable that describes the result of an endosonographic examination of a patient's oesophagus with respect to the depth of invasion of the primary tumour into the oesophageal wall; it required 80 assessments.

2.3. The patient data

For studying the ability of the oesophagus network to correctly predict the stage of a patient's cancer, the medical records of 156 patients diagnosed with oesophageal cancer are available from the Antoni van Leeuwenhoekhuis in The Netherlands. For each patient, various diagnostic symptoms and test results are available, such as the results from a gastroscopic examination of the oesophagus and an assessment of the patient's ability to swallow food. The number of data available per patient ranges between 6 and 21, with an average of 14.8. The data therefore are relatively sparse. For each patient, also the stage of his or her tumour, as established by the attending physician, is recorded. This stage can be either I, IIA, IIB, III, IVA, or IVB, in the order of advanced disease. In addition, values for various intermediate, unobservable variables are stated; these values basically are conjectures by the physician. The three most important intermediate variables pertain to the presence of haematogenous metastases, to the extent of lymph node metastases, and to the invasion of the primary tumour into the different layers of the oesophageal wall.

3. Initial experiences with probability elicitation

The oesophagus network was constructed and refined with the help of two experts in gastrointestinal oncology from The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis. In a sequence of 11 interviews of 2–4 hours each, the experts identified the relevant diagnostic and prognostic factors to be captured as statistical variables in the network, along with their possible values. The relationships between the variables were elicited from the experts using the notion of causality as a heuristic guiding principle: typical questions asked by the elicitors during the interviews were "What could cause this

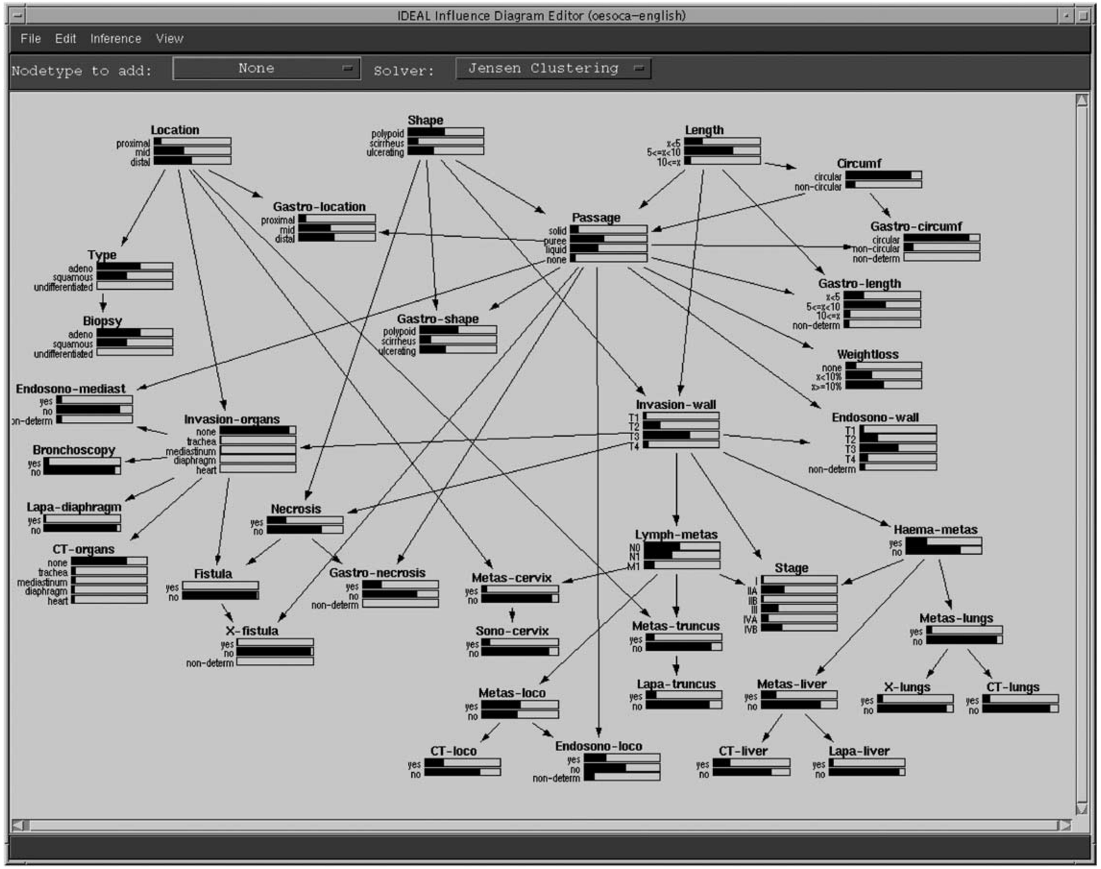


Fig. 1. The part of the oesophagus network pertaining to the cancer’s stage.

effect?'' and ''What manifestations could this cause have?''. The elicited causal relationships were expressed in graphical terms by taking the direction of causality for directing the links between related variables. Once the graphical structure of the network was considered robust, we focused our attention on obtaining the probabilities required for the network's quantitative part.

Probability assessment soon proved to be a major obstacle in the construction of our network. As in many domains, numerous sources of probabilistic information seemed to be readily available. We collected data from historical patient records and we performed an extensive literature review. Since The Netherlands is a low-incidence country for oesophageal cancer, we were not able to compose an up-to-date, large and rich enough data collection to allow for reliable assessment of the thousands of probabilities required for our network. After due consideration, we decided to retain the collected data for evaluation purposes. Literature review also did not result in ready-made assessments. Although the literature provided abundant probabilistic information, it seldom turned out to be directly amenable to encoding in our network. Research papers, for example, often reported conditional probabilities of the presence of symptoms given a cause, but not always the probabilities of these symptoms occurring in the absence of the cause. Both probabilities were required for our network, however. Also, conditional probabilities were often given in a direction opposite to the direction required. For example, the statement ''70% of the patients with oesophageal cancer are smokers'' specifies the probability of a patient being a smoker given that he or she is suffering from oesophageal cancer, while for the network the probability of oesophageal cancer developing in a smoker was required. Moreover, probabilities for unobservable intermediate disease states were lacking altogether. Another commonly found problem that prohibited direct use of the reported probabilistic information, related to the characteristics of the population from which the information was derived. These characteristics often were not properly specified or deviated seriously from the characteristics of the population for which the oesophagus network is being developed. Because of these and similar problems, hardly any probabilistic information reported in the literature turned out to be usable for our network. The knowledge and personal clinical experience of the two domain experts involved, therefore, was the single remaining source for obtaining the required probabilities.

In general, the role of domain experts in the construction of the quantitative part of a probabilistic network should not be underestimated. An expert's knowledge and experience can help not just in assessing the probabilities required, but also in fine-tuning probabilities obtained from other sources of information to the specifics of the domain at hand, and in verifying them within the context of the network. Notwithstanding, the problems that are typically encountered when eliciting probabilities from experts are widely known [8]. An expert's assessments may, for example, reflect various biases. Examples of frequently found biases are overestimation, where an expert consistently gives probability assessments that are higher than the true probabilities, and overconfidence, where assessments for likely events are too high and assessments for unlikely events are too low. Biases such as these are generally the result of the heuristics, or shortcuts, experts, often unconsciously, use for the assessment task. Moreover, the methods and presentation formats with which assessments are elicited can give rise to additional biases, especially if these do not closely match the experts' usual way of dealing with uncertainties.

Acknowledging the problems associated with human probability assessment, a number of methods have been developed in the field of decision analysis for the elicitation of judgemental probabilities [10,13]. These methods have been designed to avert to at least some extent the problems of bias and poor calibration. As these methods find widespread use in the construction of decision-analytic models, we decided to employ them with our domain experts for the assessment task. We focused on the use of a probability scale for marking assessments, on different presentation formats for the probabilities to be assessed, and on the use of gambles. Before commenting on our experiences with these methods, we would like to emphasise that, prior to the construction of the oesophagus network, our domain experts had little or no acquaintance with expressing their knowledge and clinical experience in terms of probabilities.

A well-known method for probability elicitation is the use of a *probability scale*. A probability scale is a horizontal or vertical line with some numerical anchors. Experts are asked to unambiguously mark this line with their assessment for a requested probability. The basic idea of the scale is to support experts in their assessment task by allowing them to think in terms of visual proportions rather than in terms of precise numbers. Probability scales are generally acknowledged to be easy to understand and use, and to take little time on the part of the experts involved.

The probability scale that we used with our domain experts, was a horizontal line with the three anchors 0, 50, and 100; the scale is reproduced in Fig. 2. We asked the experts to indicate the assessments for *all* conditional probabilities pertaining to a single variable given a single conditioning context on the *same* line. For example, for the context of a polypoid, circular oesophageal tumour of more than 10 cm, the experts were asked to mark the line with their assessments for the probabilities of the passage of solid food, of the passage of pureed food at best, of liquid food, and of no passage of food at all; the experts thus had to indicate four assessments on a single line. We chose to follow this procedure as we felt that it would allow the experts to compare and verify their assessments, thereby reducing the risk of overestimation. Contrary to expectation, the experts indicated that they felt quite uncomfortable working with the probability scale: it gave them ‘very little to go by’. The request to indicate several assessments on a single line further appeared to introduce a bias towards aesthetically distributed marks. This bias, commonly known as the *spacing effect* [13], seems to originate from people’s tendency to organise perceptual information so as to optimise visual attractiveness.

Another problem in our first elicitation efforts turned out to be that the probabilities to be assessed for the oesophagus network were communicated to the domain experts in mathematical notation. For example, the probability that an arbitrary patient with oesophageal cancer can swallow liquid food at best, given that he or she has a polypoid, circular primary tumour of more than 10 cm, was presented as:

$$\Pr(\text{Passage} = \text{liquid} \mid \text{Circumference} = \text{circular} \wedge \text{Shape} = \text{polypoid} \wedge \text{Length} > 10\,\text{cm})$$

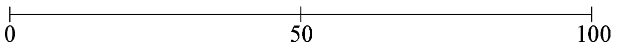


Fig. 2. The probability scale used for probability elicitation.

Our experts experienced considerable difficulty understanding conditional probabilities in this presentation format. Especially the meaning of what is represented on either side of the conditioning bar appeared to be confusing, and in fact remained to be so during successive interviews. As a result of the confusing notation, constructing a mental model of the situation referred to required considerable effort, hampering the experts focusing exclusively on the assessment task at hand.

An alternative presentation format for communicating about probabilities with experts is the *frequency format* [6]. This format builds on the observation that registering occurrences of events is a fairly automatic cognitive process requiring little conscious effort. When later asked to assess the relative frequency of the occurrence of a specific event, one may, subconsciously, review the registered events and estimate the requested frequency. The basic idea of the format therefore is to transcribe probabilities in terms of frequencies, thereby converting abstract mathematics into simple manipulations on sets of events that are easy to visualise. The frequency format generally is easier to understand for experts than mathematical notation and has been reported to be less liable to lead to biases.

For the oesophagus network, the example probability given was transcribed in the frequency format as

> Imagine 100 patients with a polypoid, circular oesophageal tumour of more than 10 cm. How many of these patients will be able to swallow liquid food at best?

Unfortunately, our experts had difficulties visualising the numbers of patients mentioned in the fragments of text: since oesophageal cancer has a low incidence in The Netherlands, visualising 100 patients with a certain combination of characteristics turned out to be a demanding, if not impossible, task.

The use of a probability scale as discussed, is a direct method for probability elicitation in the sense that experts are asked to give their assessments directly as numbers or visual proportions. With an indirect elicitation method, experts are asked not for a number or proportion but for a sequence of binary decisions from which their assessment is inferred. The use of an indirect elicitation method forestalls the need of explicitly indicating numbers and has been reported to work well for experts who do not have clear intuitions about numerical probabilities. Indirect elicitation methods are, for example, the gamble-like methods based upon the *standard reference gamble* principle [12]. The basic idea is to present an expert with a gamble, that is, a choice between two lotteries. For one of the lotteries, the probability of winning corresponds with the probability to be assessed. The probability of winning for the other lottery, termed the reference probability, is set by the elicitor. Given this explicitly set probability, the expert is asked to choose between the two lotteries. Based upon the expert's decisions, the reference probability is varied stepwise until the expert is indifferent as to which of the two lotteries is chosen. The indifference indicates that the expert judges the probability of winning to be the same for both lotteries, from which the probability to be assessed is readily inferred. Underlying the idea of a reference gamble is the assumption that people, when confronted with a gamble, try to maximise expected pay-off.

Fig. 3 shows a gamble that we used for eliciting the probabilities for our oesophagus network: the gamble pertains to the probability that an arbitrary patient with oesophageal cancer has a primary tumour of more than 10 cm in length. In the lower lottery, the elicitor

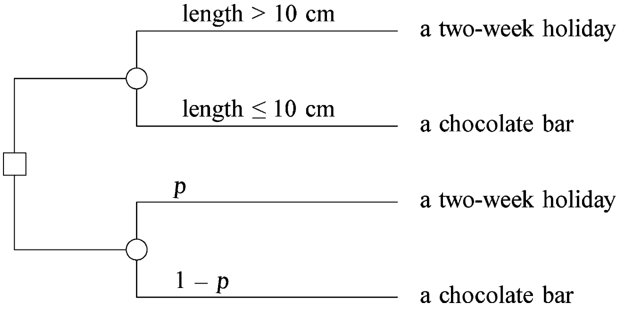


Fig. 3. An example gamble, used for elicitation of the probability of a tumour of more than 10 cm in length.

varied the reference probability p until the domain experts were indifferent between the two lotteries. The probability p' thus found equalled the experts' assessment for the probability of a tumour with a length of more than 10 cm. Unfortunately, the use of standard reference gambles with our experts was associated with several difficulties. The experts indicated that they often felt that the lotteries were very hard to conceive because of the rare or unethical situations they represented. In fact, gambling appeared to be rather demanding on the experts. Apparently, it deviated substantially from their usual cognitive processes.

Our experiences with the standard methods, from the field of decision analysis, for the elicitation of judgemental probabilities were thus unexpectedly negative. Many of the difficulties we encountered can probably be attributed to our experts' inexperience with assessing probabilities. In fact, we feel that a more extensive training would have helped to forestall at least some of these problems. Notwithstanding, we noticed that using the more involved methods especially tended to take considerable time with every single assessment. In fact, it became apparent that, even with extensive training, the elicitation of the several thousands of conditional probabilities required for our network with these methods was infeasible.

4. A method for effective probability elicitation

For the oesophagus network, several thousands of conditional probabilities had to be assessed. As we have argued in the previous section, these probabilities had to be elicited from the domain experts involved in the construction of the network. Experiences with well-known methods for probability elicitation had shown that assessing all probabilities required was not an easy task. Our negative experiences in fact induced us to design a new method for eliciting probabilities from domain experts that would enable us to elicit a large number of conditional probabilities in reasonable time.

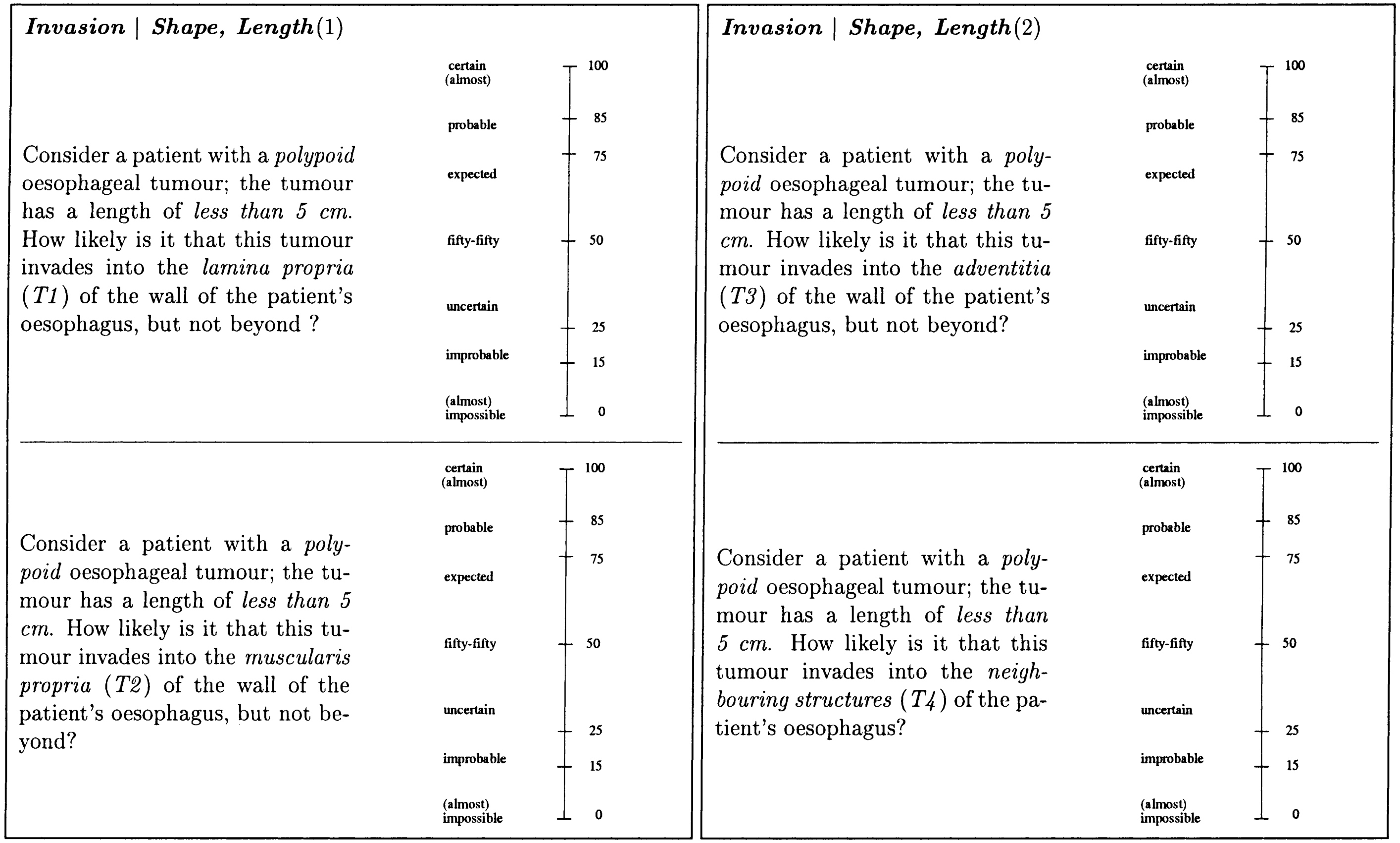
Our new method for probability elicitation from domain experts combines several different ideas. Although some of these ideas were presented before by others, we combined and enhanced them to yield a novel elicitation method. The two most important ingredients of our method are the presentation format for the probabilities to be assessed and the response scale. In communicating a conditional probability to our domain experts, we do not use mathematical notation, but instead transcribe the requested probability by a fragment of text. For the oesophagus network, for example, the probability that a patient's primary tumour invades the muscularis propria of the oesophageal wall given that the tumour is polypoid in shape and less than 5 cm in length, is presented as

> Consider a patient with a *polypoid* oesophageal tumour; the tumour has a length of *less than* 5 cm. How likely is it that this tumour invades the *muscularis propria* ($T2$) of the wall of the patient's oesophagus, but not beyond?

The fragments of text are stated in terms of *likelihood* rather than in terms of frequency to prevent difficulties with the assessment of conditional probabilities for which the conditioning context is quite rare. To support the experts in their assessment task, a response scale is depicted to the right of the text fragment. The scale in essence is a vertical line. Indicated on this line are several numerical and verbal anchors. The line is divided into six, unequally spaced, segments by the seven verbal anchors "(almost) certain", "probable", "expected", "fifty-fifty", "uncertain", "improbable", and "(almost) impossible"; on the right side of the line are the numbers 100, 85, 75, 50, 25, 15, and 0. We will presently comment on these anchors. The intended use of our scale is similar to that of the more standard numerical probability scale, that is, experts are asked to unambiguously indicate their assessment for a specific probability on the line of the scale.

The fragments of text, with the associated response scales, are grouped in such a way that the probabilities from the same conditional distribution can be taken into consideration simultaneously: they are presented in groups of two or three per page. If necessary, the various groups pertaining to the same distribution are depicted on consecutive single-sided sheets of paper so that they can be spread out on the table in front of the experts. An example is shown in Fig. 4. Explicitly grouping related probabilities has the advantage of reducing the number of times a mental switch of conditioning context is required of the domain experts during the elicitation. It also allows experts to check the coherence of their judgements.

The verbal–numerical response scale used with our method is the result of a study into the use of verbal probability expressions in dealing with uncertainty [11]. Research on human probability judgement has indicated that most people in most situations feel more at ease with verbal expressions than with numerical expressions of probability. Physicians more in specific, tend to express and process probabilities in verbal rather than numerical form. They rarely reason using numerical probabilities, and if they do, they tend to make errors [9]. Verbal probability expressions are considered to be more natural, easier to understand and communicate, and better suited to convey the vagueness of beliefs than numerical probabilities [14]. Yet, the interpretation of verbally expressed probabilities has been found to be more dependent on the context in which they are framed [2]. Also, the interpretation has been found to lead to greater within and between subject variability [3]. As there are arguments for and against the use of words and numbers, we decided to investigate the possibility of developing a scale that would support both modes of probability expression by providing verbal as well as numerical anchors. Such a double scale would allow domain experts to use either the numerical or the verbal anchors to guide them in their assessment task, the mode depending on the context and their preference.



Invasion* | *Shape, Length(1)

Consider a patient with a *polypoid* oesophageal tumour; the tumour has a length of *less than 5 cm.* How likely is it that this tumour invades into the *lamina propria* (*T1*) of the wall of the patient's oesophagus, but not beyond ?

Consider a patient with a *polypoid* oesophageal tumour; the tumour has a length of *less than 5 cm.* How likely is it that this tumour invades into the *muscularis propria* (*T2*) of the wall of the patient's oesophagus, but not beyond?

Invasion* | *Shape, Length(2)

Consider a patient with a *polypoid* oesophageal tumour; the tumour has a length of *less than 5 cm.* How likely is it that this tumour invades into the *adventitia* (*T3*) of the wall of the patient's oesophagus, but not beyond?

Consider a patient with a *polypoid* oesophageal tumour; the tumour has a length of *less than 5 cm.* How likely is it that this tumour invades into the *neighbouring structures* (*T4*) of the patient's oesophagus?

Fig. 4. Two pages with the figures pertaining to the conditional probability distribution for *Invasion*, given a polypoid oesophageal tumour with a length of less than 5 cm.

To develop a scale of verbal probability expressions to be used with numbers, we undertook four separate studies. In the first study, we asked subjects to provide a list of the verbal probability expressions they commonly use. This study yielded seven most frequently used expressions, being (translated from the corresponding Dutch expressions) "certain", "probable", "expected", "fifty-fifty", "uncertain", "improbable", and "impossible". In the second study, (other) subjects were asked to rank order these expressions. The results from this study indicated that the seven verbal probability expressions had a considerably stable rank ordering between subjects. To establish the relative distances between the seven expressions, in the third study, subjects were asked to compare each pair of expressions and assess the degree to which the two expressions conveyed the same probability. The distances generated in this study were used to project the verbal probability expressions onto a numerical scale. The expression "certain" was fixed at 100% and "impossible" was fixed at a 0% probability. The expression "probable" was calculated to be equivalent to approximately 85%, and "expected" to approximately 75%; "fifty-fifty" was calculated to be equal to 50%, "uncertain" to approximately 25%, and "improbable" to approximately 15%. Using this projection of verbal probability expressions onto numbers, the fourth study focused on the question whether decisions were influenced by the mode in which probability information was presented. The results indicated that a difference in presentation mode, that is, either verbal or numerical, did not affect our subjects' decisions. We would like to note that the four studies included subjects as well as examples from the field of medicine. For further details of the studies, we refer the reader to [11]. Since the studies had not been designed to generate verbal to numerical translations or vice versa, the verbal probability expressions could not be taken to be translations of the numerical probabilities just like that. We therefore decided to position the verbal anchors close by rather than simply beside the numerical anchors. We further decided to add the moderator "(almost)" to the extreme verbal expressions to indicate the positions of very small and very large probabilities. The resulting response scale is reproduced in Fig. 5.

As our new method was designed for the elicitation of a large number of probabilities from domain experts in little time, the probabilities obtained with the method are likely to be inaccurate and may require further fine-tuning. We therefore envision the use of our elicitation method as the first step of an elicitation procedure in which, alternately, *sensitivity analyses* are performed and probability assessments are refined. The basic idea of performing a sensitivity analysis of a probabilistic network is to systematically vary the assessments for the network's conditional probabilities and study the effects on its behaviour. Some probabilities are likely to show a considerable effect, while others will reveal hardly any influence. For the less influential probabilities, the initial assessments may suffice. For the more influential probabilities, however, refinement may be worthwhile. For example, more elaborate elicitation methods may be applied to obtain more accurate assessments for these probabilities. Given the limited and costly time of experts, it is opportune to be able to thus focus on the probabilities to which the network's behaviour shows the highest sensitivity. Iteratively performing sensitivity analyses and refining probabilities is pursued until satisfactory behaviour of the network is obtained, until the costs of further elicitation outweigh the benefits of higher accuracy, or until higher accuracy can no longer be attained due to lack of knowledge. For further information about the overall elicitation procedure envisioned, we refer the reader to [4].

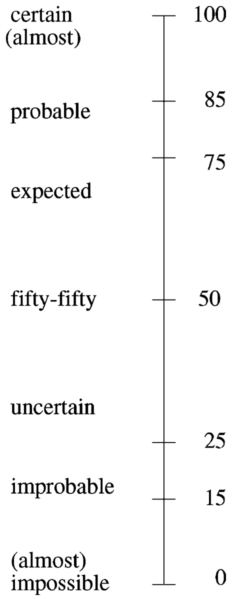


Fig. 5. The response scale with both verbal and numerical anchors.

5. Experiences with the elicitation method

We used our newly designed method for eliciting probabilities from domain experts in the construction of the quantitative part of the oesophagus network. In this section, we describe our experiences with the method. More specifically, we comment upon the observations made by our domain experts.

5.1. Using the method in practice

In the first interview with the two domain experts, we informed them of the basic ideas underlying the new elicitation method. The general format of the fragments of text was demonstrated and the intended use of the response scale was detailed. We explained the way in which the fragments of text and associated scales were grouped, and instructed the experts to take the probabilities from the same conditional probability distribution into consideration simultaneously by spreading out on the table in front of them the various sheets of paper pertaining to these probabilities. Finally, we explained to the experts that their probability assessments would be subjected to an analysis that would reveal the sensitivity of the network's behaviour to these assessments, and that, if necessary, we would try to refine the most influential ones later on. The basic idea of sensitivity analysis was explained to reassure the experts that rough assessments for the requested conditional probabilities would suffice at this stage in the construction of the network.

The elicitation of all probabilities required for the part of the oesophagus network outlined in Section 2, took five interviews of approximately two hours each. Each interview focused on a small coherent part of the network. During the interviews, the two experts jointly assessed the required probabilities. They discussed the situations described in the fragments of text and their likelihood, correcting and refining one another, before marking the response scale with their mutually agreed assessment. Prior to each interview, the elicitors spent some 10 hours preparing the fragments of text and associated response scales to be presented to the experts; after the interview, it took the elicitors 2–5 hours to process the obtained information. The new method allowed the domain experts to give their assessments at a rate of 150–175 probabilities per hour.

In the last interview, the domain experts were asked to evaluate the use of our new method of probability elicitation. For this purpose, we prepared a written evaluation form so as not to influence their observations. The domain experts were asked whether or not the different ingredients in the method had helped them in the assessment task. Also, we asked for their opinion of the specific anchors used on the response scale. The experts indicated that, overall, they had felt very comfortable with the method. They found the method most effective and much easier to use than any method for probability elicitation they had been subjected to before. Before commenting on their observations in more detail, we would like to point out that during the earlier, unsuccessful elicitation efforts, our domain experts had acquired some proficiency in expressing their knowledge and personal clinical experience in probabilities. As a result, they now appeared less daunted by the assessment task.

We recall from Section 4 that one of the ideas underlying our elicitation method is the use of a fragment of text, stated in terms of likelihood, to communicate a conditional probability to be assessed to the domain experts. During the interviews the elicitors had noticed that these fragments of text worked very well, as additional explanation of the requested probabilities was seldom necessary. The two domain experts confirmed this observation and indicated that they had had no difficulties understanding the described probabilities. The elicitors had further noted that the characteristics described in the fragments of text served to call to mind specific patients or cases from scientific papers. Although the experts could not visualise a large group of patients with certain specific characteristics, their extensive clinical experience with cancer patients in general and their knowledge of cancer growth, along with information recalled from literature, enabled them to provide the required assessments without much difficulty.

With respect to the response scale used, the domain experts indicated that they had found the presence of both numerical and verbal anchors quite helpful. They mentioned that, when thinking about a probability to be assessed, they had used words as well as numbers. Depending on how familiar they felt with the characteristics described in the fragment of text, they preferred using the verbal or numerical expressions on the scale to guide them in their assessment task. For example, the more uncertain they were about the probability to be assessed, the more they were inclined to think in terms of words. The verbal anchors of the scale then helped them to determine the position that they felt expressed the assessment they had in mind. The elicitors noticed in the consecutive interviews that it became progressively easier for the experts to express their assessments as numbers. In the first few interviews they often stated a verbal expression and then encircled the corresponding anchor or put a mark close to the anchor on the scale. In the later interviews, they considered the entire response scale, marked the scale with their assessment, and subsequently wrote a number next to their mark.

The two domain experts further mentioned that they had felt comfortable with the specific verbal anchors used on the response scale. They indicated, however, that the expression "impossible" is hardly ever used in oncology. Especially in their communication with patients, oncologists seem to prefer the more cautious expression "improbable" to refer to almost impossible events. As a consequence, our domain experts tended to interpret the expression "improbable" as a 5% or even smaller probability rather than as a probability of around 15%. However, since the response scale provided both words and numbers, they had no difficulty indicating what they meant to express. The experts also mentioned that an extra anchor for 40% would have been useful. We would like to add to these observations that our response scale hardly accommodates for indicating rather extreme probability assessments, that is, assessments very close to 0 or 100%. There are no anchors close to 0 and 100% probability on the scale since only very few subjects in our study had generated rather extreme verbal expressions. The domain experts never seemed to want to express such assessments either. When asked about this, they confirmed the correctness of our observation.

Another ingredient of our method is the grouping of the fragments of text in such a way that the probabilities from the same conditional distribution can be taken into consideration simultaneously. As mentioned before, the domain experts were advised to spread out on the table in front of them the various sheets of paper pertaining to these probabilities. They were encouraged to focus first on the probabilities from a conditional distribution that were the easiest to assess, and then to use these as anchors for distributing the remaining probability mass over the more difficult ones. This turned out to be a most effective heuristic for eliciting assessments for variables with more than two or three values. Especially in the later interviews, the domain experts were able to verify the coherence of their assessments without help and adjusted them whenever they thought fit.

5.2. The use of trends

During the elicitation interviews with our domain experts, the concept of *trend* emerged. We use the term 'trend' to denote a fixed relation between two conditional probability distributions. To illustrate the concept of trend, we address the variable *Invasion* from the oesophagus network that models the depth of invasion of the primary tumour into the wall of a patient's oesophagus. This variable can take one of the values $T1$, $T2$, $T3$, and $T4$; the higher the number indicated in the value, the deeper the tumour has invaded into the oesophageal wall and the worse the prognosis for the patient is. For the variable *Invasion*, several conditional probabilities were required, pertaining to different shapes and varying lengths of the primary tumour. Upon assessing these probabilities, the domain experts started with the probabilities for the depth of invasion of a polypoid oesophageal tumour with a length of less than 5 cm. They subsequently indicated that patients with ulcerating tumours of this length were 10% worse off with regard to the depth of invasion of the primary tumour than patients with similar polypoid tumours. They thus explicitly related the two conditional probability distributions to one another. As trends appeared to be a quite natural way of expressing probabilistic information, we encouraged the experts to provide trends wherever appropriate.

We designed a generic method for dealing, in an intuitively appealing and mathematically sound way, with the trends provided by our domain experts. The method is best

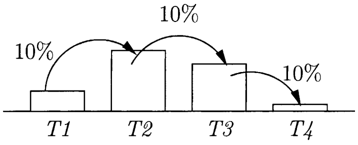


Fig. 6. A schematic representation of handling trends.

explained in terms of the example trend given. Suppose that, given a polypoid oesophageal tumour of less than 5 cm in length, the probabilities for the four different values of the variable *Invasion* are assessed at x_1, x_2, x_3, and x_4—x_i being the probability assessment for the value *Ti*. The probabilities x_i, $i = 1, \ldots, 4$, constitute the *anchor distribution* that is to be adjusted by the indicated trend to compute the probabilities for the related distribution. After close consultation with our domain experts, we interpreted the specified trend as stating that 10% of the patients with a polypoid tumour of less than 5 cm with *Ti* for its depth of invasion would have had *Ti* + 1 for the depth of invasion if the tumour were an ulcerating tumour, $i = 1, \ldots, 3$. The basic idea of the interpretation of the trend is depicted in Fig. 6. For the probability assessments $y_1, \ldots, y_4$ for the different values of the variable *Invasion* given an ulcerating tumour of less than 5 cm, we now define

$$\begin{aligned}
y_1 &\leftarrow x_1 - 0.10x_1\\
y_2 &\leftarrow x_2 - 0.10x_2 + 0.10x_1\\
y_3 &\leftarrow x_3 - 0.10x_3 + 0.10x_2\\
y_4 &\leftarrow x_4 + 0.10x_3
\end{aligned}$$

It is readily verified that $y_1, \ldots, y_4$ lie between 0 and 1, and together sum up to 1. In addition, it will be evident that this method of handling trends can easily be generalised to variables with an arbitrary number of values and to trends specifying other percentages and other directions of adjustment.

6. A preliminary evaluation study

To gain insight in the quality of the probabilities obtained with our new elicitation method, we conducted a preliminary evaluation study of the oesophagus network. In this study, we used data from patients from the Antoni van Leeuwenhoekhuis diagnosed with oesophageal cancer. In Section 6.1, we reflect on the probabilities obtained; we compare them against the available data in Section 6.2. In Section 6.3, we focus on the elicited probabilities in the context of the network. For this purpose, we entered, for each patient, all diagnostic symptoms and test results available and computed the most likely stage of the patient's cancer from the network; we subsequently compared the computed stage against the stage recorded in the data.

6.1. The obtained probabilities

The part of the oesophagus network outlined in Section 2, includes 39 statistical variables. For these variables, 900 probabilities were required, constituting a total of

267 (conditional) probability distributions. The number of probabilities to be assessed per variable ranged between 3 and 144.

Many of the assessments that we obtained from our domain experts, equalled either 0 or 1: the experts gave 312 zeroes and 100 ones, together amounting to 46% of the network's probabilities. We would like to note that 144 of these probabilities pertain to the deterministic variable that models the cancer's stage, that is, 35% of the zeroes and ones serve to constitute the conditional probability distributions for a *single* variable. The domain experts further specified many probabilities on the lower half of the response scale: 72% of their assessments were less than or equal to 0.50. For 12 of the 39 variables in the network, the domain experts indicated trends, as described in the previous section. Using these trends, 241 probabilities were computed from other assessments. Of the total of 900 probabilities, therefore, 73% were assessed directly and 27% indirectly by adjustment of other probabilities. The indirect assessments pertained to 65 different conditional probability distributions. The trends indicated by the domain experts ranged from equal to the anchor distribution to a 20% adjustment, in either direction, from this distribution.

To study the overall distribution of the assessments obtained with our elicitation method, we performed a frequency count over the network's quantitative part. Fig. 7(a) summarises the frequencies of all assessments obtained, be it directly or indirectly; we restricted the figure to the assessments that are not equal to zero or one to enhance discernibility of the other frequency counts. Fig. 7(b) shows the frequencies of the assessments that were specified directly by the domain experts; once again we excluded zero and one from the figure.

We recall from Section 4 that the response scale used with our elicitation method specifies seven numerical anchors: 0, 15, 25, 50, 75, 85, and 100, or, alternatively, 0, 0.15, 0.25, 0.50, 0.75, 0.85, and 1.00. By comparing our experts' assessments with these anchors, we find that 54% of all assessments and 63% of all direct assessments coincide with anchors. Focusing on the non-extreme assessments, that is, excluding 0 and 1.00, we find that 16% of all assessments and 20% of the direct assessments are anchors. The frequency counts further reveal that among the 10 most often specified assessments, there are four anchors from the response scale: 0, 0.15, 0.85, and 1.00. Among the 10 most frequently specified direct assessments, there are even six anchors: 0, 0.15, 0.25, 0.75, 0.85, and 1.00. These findings are consistent with the often reported observation that the external stimulus used, in our case the response scale, plays a dominant role in the elicitation process. If anchors are presented, assessors are more inclined to use these than place their marks in between anchors, thereby possibly introducing a bias towards the anchors. Our findings suggest that such a bias may be present in the assessments that we obtained for the oesophagus network.

To conclude our discussion of the probabilities obtained, we observe that, while the experts indicated that an extra anchor for 0.40 would have been helpful, they gave this assessment only seven times.

6.2. A comparison against the data

As described in Section 3, we had not been able to compose a large and rich enough data collection to allow for reliable assessment of the probabilities required for the oesophagus

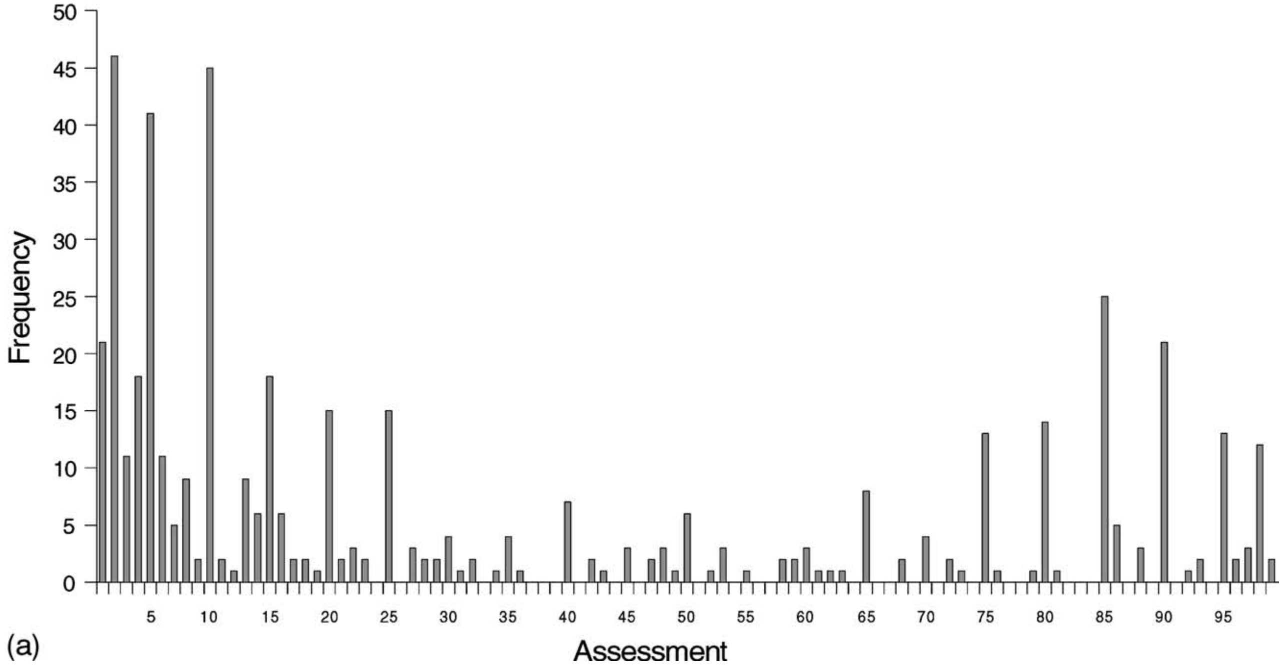


Fig. 7. The distribution of all assessments obtained (a), and of the assessments that were specified directly (b); 0 and 100% are excluded to enhance discernibility.

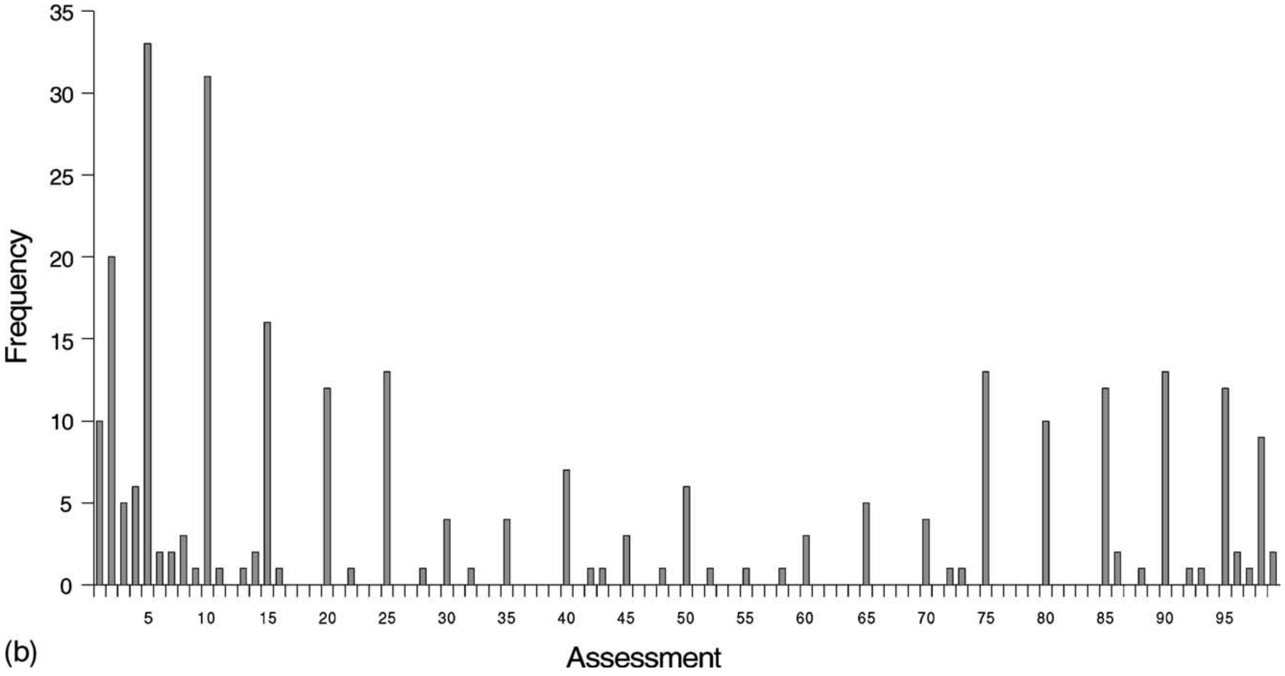


Fig. 7. (*Continued*).

network. Our efforts to compose such a collection, however, had resulted in data from the historical records of 156 patients diagnosed with oesophageal cancer from the Antoni van Leeuwenhoekhuis, as outlined in Section 2. In this section, we compare the probabilities given by our domain experts against estimates computed from these data. Before doing so, we would like to note that the data collection does not constitute a fully independent source of information, as the collection consists of data from patients treated by our domain experts and their colleagues at the Antoni van Leeuwenhoekhuis. However, since the historical records date back to between 1978 and 1985, and at that time the experts had not been actively assessing probabilities as we are asking them to do now, it is almost impossible that they would retrieve probability assessments for the data collection from memory. They may, of course, remember the patients they had treated before, but not at such a level of detail that the data would be too much dependent to render the comparison less meaningful.

We estimated, from our data collection, as many probabilities for the oesophagus network as possible. For only 26 of the 39 statistical variables involved, however, probability estimates could be obtained: the remaining 13 variables were not recorded in the data. Furthermore, for the variables that were recorded, not all probabilities required could be estimated, as several combinations of values were missing from the data collection. For example, the data collection did not include any patients with a non-circular ulcerating tumour with a length of less than 5 cm: this combination of presentation characteristics is not impossible, but merely unlikely. The data all in all provided for the estimation of 368, or 41%, of the network's probabilities, pertaining to 125 conditional distributions.

To investigate whether or not the probability assessments given by our domain experts were in the same range as the estimates obtained from the data, we computed a 95%-confidenceinterval for each of the 368 probability estimates. The 95%-confidenceinterval of a specific estimate is the interval in which the 'true' probability lies with 95% certainty. The length of the interval thus quantifies the uncertainty in the estimate. For a probability estimate p, its 95%-confidenceinterval was approximated as follows:

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}\right] \cap [0, 1]$$

where n is the number of patients whose data were used in the computation of the estimate p. From the formula it is readily seen that the larger the number of patients on which the estimate is based, the smaller the estimate's 95%-confidenceinterval. The confidence intervals that we obtained for our probability estimates differed in length, due to the varying availability of data. For example, for the estimate 0.50 for the probability of an amount of weight loss of more than 10% in patients who are able to swallow liquid food at best, we found a 95%-confidenceinterval of $[0.38, 0.62]$, based on the data of 66 patients; for the estimate 0.60 for the same amount of weight loss in patients who are not able to swallow any food at all, the computed interval equalled $[0.17, 1.00]$, based on the data of just five patients. The computed confidence intervals were rather large as a result of data sparseness; we found an average length of 0.25. For 250 of the 368 estimates, the 95%-confidenceinterval covered the assessment that we had elicited from our experts. So, from the assessments that could be compared against the data, 68% were more or less similar to the computed probability estimates.

As mentioned before, our domain experts had indicated trends for 12 statistical variables, pertaining to 65 different conditional probability distributions. For 23 of these 65 trends, we could compare the probabilities from both the specified anchor distribution and the distribution computed from the anchor, against probability estimates from the data. To determine whether a specific distribution specified by the experts matched the data, we conducted a number of χ^2-tests. A χ^2-test builds upon a specific distribution for the difference between two probability distributions. This difference is measured as follows:

$$k = \sum_{x_i} \frac{(\Pr(x_i) - \widehat{\Pr}(x_i))^2}{\widehat{\Pr}(x_i)}$$

where Pr is the *observed* probability distribution, that is, the distribution estimated from the data, and $\widehat{\Pr}$ is the *expected* distribution as given by the experts; the values x_i over which the summation is performed, are the values of the statistical variable to which the two probability distributions pertain. If the probability of k is less than or equal to 5%, then the difference between the observed distribution and the estimated distribution is statistically significant, from which we then conclude that the two distributions are not sufficiently similar. Fig. 8 summarises the comparison results that we obtained for the various trends.

For 15, or 65%, of the 23 trends, the anchor distribution given by the experts did not significantly differ from the same distribution estimated from the data. For eight of these 15 trends, the probability distribution that was computed from the anchor distribution by adjustment did not significantly differ from the same distribution estimated from the data either. For 35% of the trends specified by the experts, therefore, both the anchor distribution and the computed distribution closely matched the data. Of the eight trends of which the anchor distribution given by the experts differed significantly from the distribution estimated from the data, we found for three of them that also the computed distribution did not match the data. For 13% of the trends, therefore, both the anchor distribution and the computed distribution differed significantly from the distributions estimated from the data.

For the eight trends of which both the anchor distribution and the computed distribution matched the data, we may conclude that the direction as well as the percentage of adjustment that were indicated by our domain experts were closely reflected in the data

	anchor distribution	*computed distribution*	*both distributions*
match	15	13	8
no match	8	10	3

Fig. 8. The number of matching anchor and computed distributions.

collection. For the three trends of which both the anchor distribution and the computed distribution did not match the data, we investigated whether or not the specified trend itself was reflected in the data collection. For this purpose, we applied the trend, not to the anchor distribution given by the experts, but to the same distribution estimated from the data. For one of these trends, the thus computed probability distribution closely matched the data. We conclude that for a total of nine trends, that is, for 39% of the trends specified by the domain experts, the indicated direction and percentage of adjustment were reflected in the data collection. Alternatively, 61% of the trends appeared not to be present in the data. Upon examining the 14 apparently mismatching trends, we found that for four of them a related trend seemed to be present in the data collection: for either an opposite direction or a weaker percentage of adjustment, the computed distribution matched the data. We would like to note that for many of the trends given by our experts only very few patient data were available as a basis for comparison. As a consequence, we feel that no conclusive statements with regard to the specified trends can be made.

6.3. The quality of the network

To conclude our preliminary evaluation of the elicited probabilities, we studied the performance of the oesophagus network with the available patient data. The study again focused on the part of the network that provides for establishing the stage of a patient's cancer; we recall that this stage can be either I, IIA, IIB, III, IVA, or IVB, in the order of progressive disease.

In a first study of our network, we entered, for each patient from the data collection, all diagnostic symptoms and test results available. We then computed the most likely stage of the patient's cancer from the network and compared it against the stage recorded in the data. Fig. 9 shows the results from this first study. For 80 of the 156 patients, the stage of the cancer recorded in the data matched the stage that was computed from the network to have the highest probability. Assuming that the stages recorded in the data are correct, we conclude that the network established the correct stage for 51% of the patients. We would like to note that it is not uncommon to find a percentage in this range in initial evaluations of knowledge-based systems [1].

In trying to identify the reasons for the network's relatively poor performance, we carefully examined the patient data. In doing so, we identified three major problems. For 10 patients, the stage recorded in the data was acknowledged by the domain experts to be incorrect on retrospection. Various other anomalies in the data constituted the second problem. For example, for some patients a deeper invasion of the primary tumour into the oesophageal wall was found during surgery than conjectured from endosonographic findings. For these patients, the pre-surgical findings and the post-surgical stage were recorded in the data. Because only the (pre-surgical) findings had been entered, the network had yielded a stage different from the recorded one. The third major problem was found in the way that findings had been entered into the patients' medical records. Often no distinction was made between facts and findings from diagnostic tests. For example, for many patients the medical record stated the presence or absence of lymphatic metastases near the truncus coeliacus without indicating how this fact had been established. Without explicitly stated test results, the network could not establish the presence or absence of

		network						
		I	IIA	IIB	III	IVA	IVB	*total*
	I	**2**	0	0	0	0	0	2
	IIA	0	**34**	0	3	0	0	37
data	IIB	0	3	**0**	3	0	0	6
	III	1	16	1	**24**	1	1	44
	IVA	1	9	2	23	**6**	1	42
	IVB	0	2	0	8	1	**14**	25
	total	4	64	3	61	8	16	156

Fig. 9. The results from the first study.

these metastases, which often resulted in an incorrect stage. The network so far included a single diagnostic test for establishing the presence or absence of metastases near the truncus coeliacus. This diagnostic test, a laparoscopic procedure, is rather invasive and has only recently been introduced into clinical practice. As it is very unlikely that this test had been performed in the majority of the patients from our data collection, we concluded that some variables modelling diagnostic tests were missing from our network.

Building upon the aforementioned observations, we decided to conduct a second study of the oesophagus network. For this purpose, we corrected the erroneous stages and the other anomalies in the data, that is, as far as they had been identified by our experts. As we carefully examined and, if necessary, corrected the data of every single patient regardless of whether or not the network had established the correct stage, we felt that the results from the first study would not bias the results of this second study. For the second study, we extended the network with three extra statistical variables. These variables model two additional diagnostic tests for establishing the presence or absence of metastases in the lymph nodes near the truncus coeliacus and one test for establishing the presence or absence of lymphatic metastases in the neck. We entered for each patient the available symptoms and test results, as before. If no diagnostic tests were specified explicitly for facts with regard to lymphatic metastases in the neck or near the truncus coeliacus, we entered these facts as test results for the appropriate newly included variables. In addition, we entered for each patient the facts stated in the data for which an indication of the test performed was missing; on average, 0.4 additional facts were entered per patient. The overall results of the second study are shown in Fig. 10.

The figure reveals that for 132 of the 156 patients, the stage of the cancer as recorded in the (modified) data matched the stage computed from the network. Again assuming that the stages recorded in the data are correct, the network now established the correct stage for 85% of the patients.

		network						
		I	IIA	IIB	III	IVA	IVB	*total*
	I	**2**	0	0	0	0	0	2
	IIA	0	**37**	0	1	0	0	38
data	IIB	0	1	**0**	3	0	0	4
	III	1	11	0	**35**	0	0	47
	IVA	0	0	0	4	**35**	0	39
	IVB	0	0	0	3	0	**23**	26
	total	3	49	0	46	35	23	156

Fig. 10. The results from the second evaluation.

7. Concluding observations

With the help of two experts in gastrointestinal oncology from The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, a decision-support system is being developed for patient-specific therapy selection for oesophageal cancer. The kernel of the system is a probabilistic network that describes the presentation characteristics of a tumour of the oesophagus and the pathophysiological processes of its invasion and metastasis; in addition, it describes the possible effects of the available treatment alternatives. In the construction of our network, we found that probability elicitation can be a major obstacle. Building upon negative experiences with existing methods, we designed a new method for eliciting probabilities from domain experts that allows for the elicitation of large numbers of probabilities in reasonable time. Our elicitation method combines several ideas, among which are the ideas of transcribing probabilities as fragments of text and of using a response scale with both numerical and verbal anchors. We used our new elicitation method for obtaining the probabilities required for a coherent and self-contained part of the oesophagus network. Our domain experts indicated that they found the method much easier to use than any method for probability elicitation they had been subjected to before. Moreover, the method allowed the domain experts to give their assessments at a rate of over 150 probabilities per hour. So far our response scale has only been used by the two experts involved in the construction of the oesophagus network. To establish whether our scale would be helpful to other experts in other settings as well, we are currently conducting a comprehensive study into its usability with a large number of subjects with different backgrounds. Preliminary results suggest that our verbal–numerical scale improves the assessment process when compared against the more conventional numerical probability scale.

To gain some insight in the quality of the probabilities obtained with our new elicitation method, we conducted a preliminary evaluation study of the oesophagus network. After correcting various anomalies in the data, we found that a correct stage was established by the network for 85% of the patients. Given that the probabilities used are rough initial assessments and that the patient data require further cleaning up, the results from this preliminary study are quite encouraging. Before any conclusive statements about the quality of the probabilities can be made, however, more extensive evaluation studies of our network are required.

For the construction of the oesophagus network, our newly designed elicitation method meant a major breakthrough. Prior to the use of our method, we had spent over a year experimenting, on and off, with other methods for probability elicitation, without success. Using our elicitation method, the probabilities for a major part of the oesophagus network were elicited in reasonable time. Our method seems to us to be well suited for eliciting the large number of probabilities that are typically required for a realistic probabilistic network. Although our method tends to require considerable time from the elicitors for preparing for the interviews with the experts, we feel that the ease with which probabilities are subsequently elicited makes this time well spent.

References

[1] Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med 1994;330:1792–6.

[2] Brun W, Teigen KH. Verbal probabilities: ambiguous, context-dependent, or both? Organizational Behav Hum Decision Processes 1988;41:390–404.

[3] Budescu DV, Weinberg S, Wallsten TS. Decisions based on numerically and verbally expressed uncertainties. J Exp Psychol: Hum Percept Performance 1988;14:281–94.

[4] Coupé VMH, van der Gaag LC, Habbema JDF. Sensitivity analysis: an aid for belief-network quantification. Knowledge Eng Rev 2000;15:1–18.

[5] Druzdzel MJ, van der Gaag LC. Building probabilistic networks: where do the numbers come from? IEEE Trans Knowledge Data Eng 2000;12:481–6.

[6] Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. Psychol Rev 1995;102:684–704.

[7] Jensen FV. Bayesian Networks and Decision Graphs. New York: Springer; 2001.

[8] Kahneman D, Slovic P, Tversky A. Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press; 1982.

[9] Kuipers B, Moskowitz AJ, Kassirer JP. Critical decisions under uncertainty: representation and structure. Cognitive Sci 1988;12:177–210.

[10] Morgan MG, Henrion M. Uncertainty, A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge: Cambridge University Press; 1990.

[11] Renooij S, Witteman CLM. Talking probabilities: communicating probabilistic information with words and numbers. Int J Approximate Reasoning 1999;22:169–94.

[12] von Neumann J, Morgenstern D. The Theory of Games and Economic Behavior. 3rd ed. New York: Wiley; 1953.

[13] von Winterfeldt D, Edwards W. Decision Analysis and Behavioral Research. Cambridge: Cambridge University Press; 1986.

[14] Wallsten TS, Budescu DV, Zwick R. Comparing the calibration and coherence of numerical and verbal probability judgments. Manage Sci 1993;39:176–90.