Probabilities for a Probabilistic Network: A Case-study in Oesophageal Carcinoma

L.C. van der Gaag, S. Renooij, C.L.M. Witteman,
Utrecht University, Institute of Information and Computing Sciences,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
{linda,silja,cilia}@cs.uu.nl

B.M.P. Aleman, and *B.G. Taal*
The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis,
Department of Radiation Oncology and Gastroenterology,
Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
{baleman,bgtaal}@nki.nl

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 carcinoma. The kernel of the system is a probabilistic network that describes the characteristics of oesophageal carcinoma 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. 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. The method allowed us to elicit the many probabilities required for our network in little time. Using data from 185 patients, we conducted an evaluation study to assess the quality of the probabilities obtained. We found that for 85% of the patients, our probabilistic network yielded the correct outcome.

1 Introduction

The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, is a specialised center for the treatment of cancer patients. Every year some eighty patients receive treatment for *oesophageal carcinoma* at the center. 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 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 constructed with the help of two experts in gastrointestinal oncology from the Netherlands Cancer Institute, who are the co-authors B.M.P. Aleman and B.G. Taal of the present paper. The system is destined for use in clinical practice.

The kernel of our decision-support system is a *probabilistic network*. A probabilistic network is a 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 [Jensen, 1996]. The probabilistic network in our decision-support system models various characteristics of an oesophageal carcinoma, 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 a carcinoma's stage. For prognostication, the network in addition describes the possible effects of the available therapeutic alternatives. When a patient's symptoms and test results are entered, the network provides for establishing the stage of the patient's carcinoma and for predicting 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 carcinoma.

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 the characteristics of an oesophageal carcinoma and the possible effects of the different therapies available. We then focused on the elicitation of the probabilities required for the quantitative part of the network. The task of eliciting probabilities is generally acknowledged to be the most daunting in constructing a probabilistic network [Druzdzel & Van der Gaag, 2000]. In the domain of oesophageal carcinoma, various sources of probabilistic information appeared to be readily available for the task. However, neither data collection nor a thorough literature review yielded any usable results. The single remaining source of probabilistic information, therefore, was the knowledge and personal clinical experience of the two domain experts involved in the project.

Various methods for eliciting judgemental probabilities from experts are available from the field of decision analysis, ranging from probability scales for marking assessments to gambles [Morgan & Henrion, 1990, Von Winterfeldt & Edwards, 1986]. For eliciting the probabilities required for the oesophagus network, we set out using these well-known methods with our domain experts. 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 soon 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 required for a moderately sized probabilistic network.

Building upon our negative experiences with existing methods, we designed a new method for eliciting probabilities from domain experts. We tailored our method to the elicitation of a large number of probabilities in little time. Our method combines several ideas, such as *transcribing* the probabilities to be assessed as fragments of text and providing a scale with both *numerical* and *verbal* anchors for marking assessments. Using our method in the construction of the oesophagus network, our domain experts provided the probabilities required at a rate of over 150 numbers per hour.

To assess the quality of the probabilities obtained with our new elicitation method, we conducted an evaluation study of the oesophagus network, using data, from the Antoni van Leeuwenhoekhuis, from 185 patients diagnosed with oesophageal carcinoma. The evaluation study focused on the part of the network that provides for establishing the *stage* of a patient's carcinoma. This stage summarises the carcinoma's characteristics, its depth of invasion, and the extent of its metastasis, and is indicative of the likely outcome of treatment. We would like to note that in our decision-support system the characteristics, depth of invasion, and extent of metastasis themselves are of interest rather than the stage derived from them. Focusing on the summarising stage, however, provides overall insight in the diagnostic part of the network. We found that for 85% of the patients, the stage established by the network as the most likely stage matched the stage that was recorded in the patient's data.

In this paper, we describe the oesofagus network, its construction, and its evaluation. In Section 2 we give an overview of the network. In Section 3 we describe our initial experiences with probability elicitation. In Section 4 we detail the method that we designed for eliciting a large number of probabilities from experts. In Section 5 we evaluate the use of our method in the construction of the oesophagus network; more specifically, we comment on the observations made by our domain experts. In Section 6 we present the results of the evaluation study of the network. The paper ends with some concluding observations in Section 7.

2 The oesophagus network

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 carcinoma may develop in a patient's oesophagus. An oesophageal carcinoma has various characteristics that influence its prospective growth. These characteristics include the location of the carcinoma in the oesophagus and its histological type, its length, and its macroscopic shape. An oesophageal carcinoma 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 carcinoma may result in lymphatic metastases in distant lymph nodes and in haematogenous metastases in, for example, the lungs and the liver. The characteristics, depth of invasion, and extent of metastasis, summarised in the carcinoma's stage, largely influence a patient's life expectancy and are indicative of the effects and complications to be expected from the different available therapeutic 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 gastroscopic and endosonographic examination of the oesophagus and a CT-scan of the patient's chest and liver. These tests differ considerably in their sensitivity and specificity characteristics. For example, endosonography for establishing the presence or absence of metastases in the loco-regional lymph nodes, has a low sensitivity and specificity whereas gastroscopy for establishing the carcinoma's shape, has considerably better sensitivity and specificity characteristics.

Whereas establishing the presence of an oesophageal carcinoma in a patient is relatively straightforward, the staging of the carcinoma and especially the selection of an appropriate therapy are far harder tasks. In the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, different therapeutic alternatives are available, ranging from surgical removal of the oesophagus to positioning a prosthesis in the oesophagus. The effects aimed at by instilling a therapy include removal or reduction of the patient's primary tumour to prolong life expectancy and an improved passage of food through the oesophagus. The therapies differ in the extent to which these effects can be attained. For example, where the aim of surgical removal of the oesophagus is to achieve 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 problems with swallowing food. Instillation of a therapy is often accompanied not only 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 therapeutic alternatives for a specific patient depend on the characteristics of his or her carcinoma, on the depth of invasion of the carcinoma into the oesophageal wall and neighbouring structures, and on the extent of the carcinoma's metastasis.

We captured the state-of-the-art knowledge about oesophageal carcinoma and its treatment in a probabilistic network, also known as a Bayesian network or causal network, [Jensen, 1996]. 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 carcinoma 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 carcinoma currently includes over 70 statistical variables. More than 4000 conditional probabilities have been specified. The graphical structure and its associated probabilities uniquely capture a joint probability distribution over the represented variables. Any probability of interest can therefore be computed from the network. More specifically, the stage of a patient's carcinoma 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 carcinoma'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 carcinoma. This part constitutes a coherent and self-contained probabilistic network. The network's graphical structure is depicted in Figure 1; the figure also shows the prior marginal probability distribution for every statistical variable. The 40 variables involved required some 1000 probability assessments. The variable requiring the largest number of assessments, 144, models the stage of a carcinoma; this variable is a deterministic variable classifying an oesophageal carcinoma in one of six categories of disease. 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 carcinoma into the oesophageal wall; it requires 80 assessments.

3 Initial experiences with probability elicitation

The oesophagus network is constructed and refined with the help of two experts in gastrointestinal oncology from the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis. In a sequence of eleven interviews of two to four 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: typical questions asked by the elicitors during the interviews were "What could cause this effect ?" and "What manifestations could this cause have ?". The thus 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 the elicitation of the probabilities required.

Probability elicitation soon proved to be a major obstacle in the construction of the oesophagus 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 a literature review. Unfortunately, the Netherlands being a low-incidence country for oesophageal carcinoma, we were not able to compose an up-to-date, large and rich enough data collection to allow for reliable assessment of all probabilities required; after due consideration, we decided to save 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

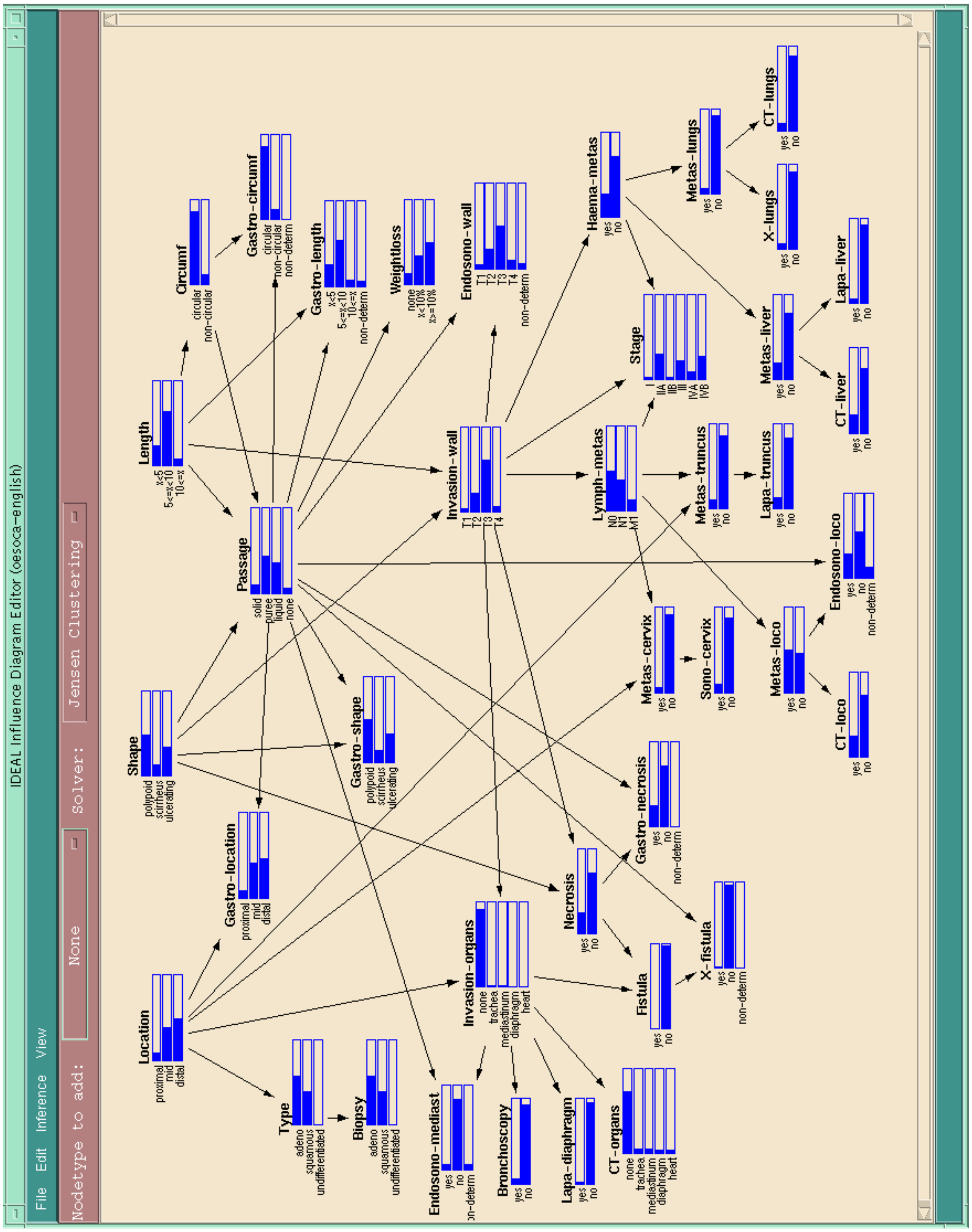

Figure 1: The part of the oesophagus network pertaining to the stage of a carcinoma.

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 results 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 of probabilistic information.

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 to the specifics of the domain at hand, and in verifying them within the context of the network. However, the problems encountered when eliciting probabilities from experts are widely known, e.g. [Kahneman *et al.*, 1982]. An expert's assessments, for example, may reflect various biases and may not be properly calibrated. Examples of 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 they do not closely match the experts' usual way of dealing with uncertainties.

Acknowledging these problems, a number of methods have been developed in the field of decision analysis for the elicitation of unbiased probabilities from experts [Morgan & Henrion, 1990, Von Winterfeldt & Edwards, 1986]. As these methods have found widespread use in the construction of decision-analytic models, we decided to employ them in our efforts to elicit probabilities for the oesophagus network. 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 numerical anchors. Experts are asked to unambiguously mark their assessment for a requested probability on this scale. 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 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 we used with our domain experts was a horizontal line with the three anchors 0, 50, and 100; the scale is shown in Figure 2. We asked the experts to mark the assessments for *all* conditional probabilities pertaining to a single variable given a single conditioning context on the same scale. For example, for the context of a polypoid, circular carcinoma of more than 10 centimeters, the experts were asked to mark their assessments for the probabilities of the passage of solid food, for the passage

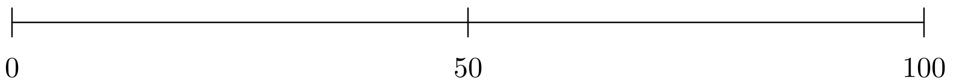

Figure 2: The probability scale used for probability elicitation.

of puréed food at best, of liquid food, and of no passage at all; the experts thus had to indicate four assessments on a single scale. We chose to follow this procedure as 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 mark several assessments on a single line further appeared to introduce a bias towards aesthetically distributed marks. This bias, commonly known as the *spacing effect* [Von Winterfeldt & Edwards, 1986], 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 carcinoma of more than 10 centimeters, was presented as

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

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. As a result, the experts had difficulties constructing a mental model of the situation referred to and could not focus exclusively on the assessment task at hand.

An alternative presentation format for communicating probabilities to experts is the *frequency format* [Gigerenzer & Hoffrage, 1995]. This format builds on the observation that registering occurrences of events is a fairly automatic cognitive process requiring little conscious effort. The basic idea is to transcribe probabilities in terms of frequencies, thereby converting abstract mathematics into simple manipulations on sets that are easy to recall and 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 above was transcribed in the frequency format as

> Imagine 100 patients with a circular, polypoid oesophageal carcinoma of more than 10 centimeters. 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 carcinoma has a low incidence in the Netherlands,

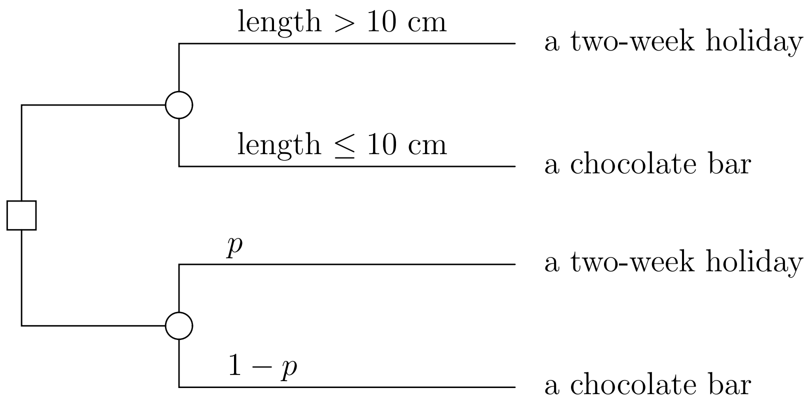

Figure 3: An example gamble, used for elicitation of the probability of a tumour of more than 10 centimeters in length.

visualising one hundred 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 above 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. For experts who do not have clear intuitions about numerical probabilities, the use of an indirect elicitation method forestalls the need of explicitly indicating numbers, especially very small ones. Indirect elicitation methods are, for example, the gamble-like methods based upon the *standard reference gamble* principle [Von Neumann & Morgenstern, 1953]. 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 is set by the elicitor. The latter probability is varied 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 this idea is the assumption that people, when confronted with a gamble, try to maximise expected pay-off.

Figure 3 shows a gamble that we used for the oesophagus network: the gamble pertains to the probability that an arbitrary patient with oesophageal carcinoma has a tumour of more than 10 centimeters in length. In the lower lottery, the elicitor varied the probability p until the domain experts were indifferent between the two lotteries, indicating that the probability of a carcinoma with a length of more than 10 centimeters equaled the indifference probability p.

Unfortunately, the use of standard reference gambles with our experts was hampered by 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 for the experts, as it did not correlate with their usual cognitive processes. Moreover, the use of lotteries tended to take so much time that it soon became apparent that the elicitation of several thousands of conditional probabilities in this way was quite 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. Experience with well-known methods for probability elicitation had shown that assessing all probabilities required was not an easy task. Our negative experiences with these methods 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 little 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 and, as we will argue in the next section, effective 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 carcinoma invades the muscularis propria of the oesophageal wall given that the carcinoma is polypoid in shape and less than 5 centimeters in length, is presented as

> Consider a patient with a *polypoid* oesophageal carcinoma; the carcinoma has a length of *less than 5 cm*. How likely is it that this carcinoma invades the *muscularis propria (T2)* of the patient's oesophageal wall, 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 a conditional probability for which the conditioning context is quite rare. To support the experts in their assessment task, a vertical response scale is depicted to the right of the text fragment. Indicated on this scale are several numerical and verbal anchors. The scale 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 scale are the numbers 100, 85, 75, 50, 25, 15, and 0. We will presently comment on the specific anchors used.

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, if necessary 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 Figure 4. Explicitly grouping related probabilities has the advantage of reducing the number of times a mental switch of conditioning context

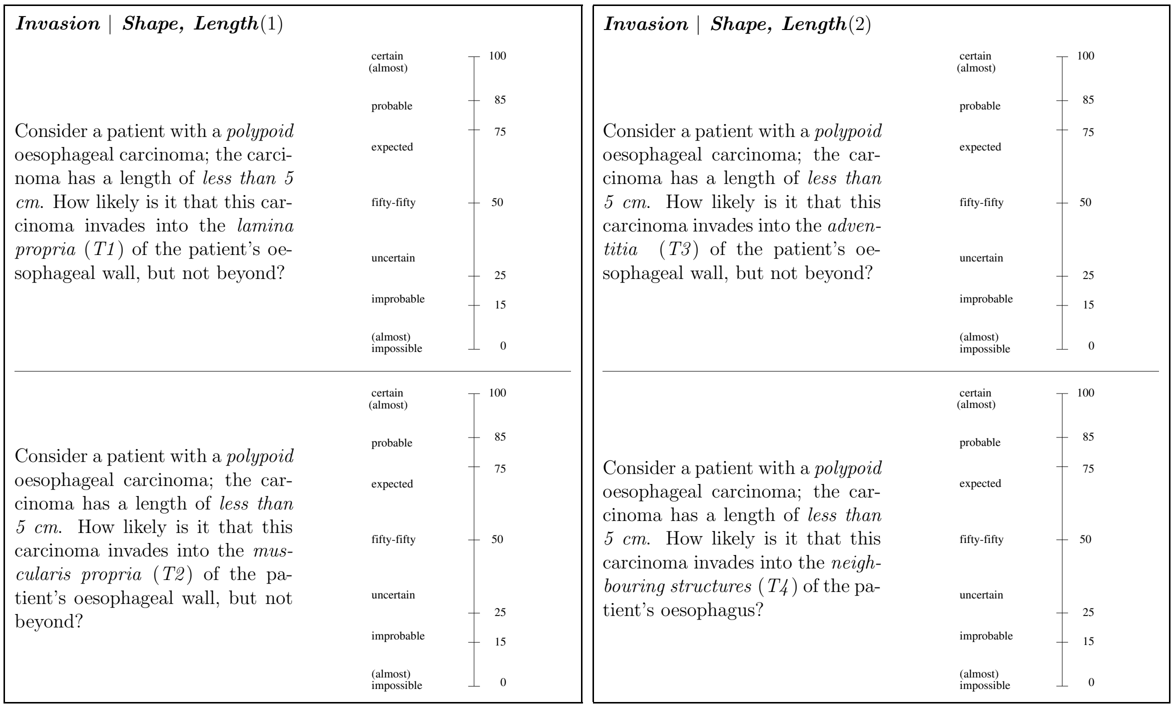

Figure 4: Two pages with the figures pertaining to the conditional probability distribution for *Invasion*, given a polypoid carcinoma with a length of less than 5 cm.

is required of the domain experts during the elicitation. It also allows experts to check the coherence of their judgments.

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. Research on human probability judgement has indicated that most people in most situations tend to feel more at ease with verbal expressions than with numerical expressions of probability. Verbal probability expressions are considered to be more natural, easier to understand and communicate, and better suited to convey the vagueness of beliefs [Wallsten *et al.*, 1993]. On the other hand, the interpretation of verbally expressed probabilities has been found to be more dependent on the context in which they are framed [Brun & Teigen, 1988]; also, the interpretation has been found to lead to greater within and between subject variability [Budescu *et al.*, 1988]. As there are arguments for and against the use of both words and numbers, we decided to investigate the possibility of developing a scale with *both* modes of probability expression, allowing subjects to use either one depending on the context and their preference.

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)

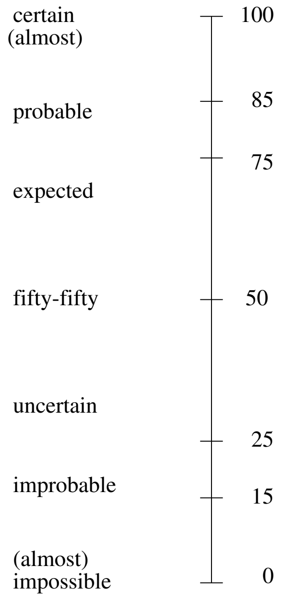

Figure 5: The response scale with both verbal and numerical anchors.

"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 an extended paper [Renooij & Witteman, 1999]. Because people may have different preferences in different situations, we decided to include both the numerical and the verbal anchors on our response scale. Since the verbal probability expressions were explicitly not intended as translations of the numerical probabilities, we decided to position them 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 Figure 5.

As our new elicitation method was designed for the elicitation of a large number of probabilities from domain experts in little time, the obtained probabilities 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 over a plausible interval and study the effects on its behaviour. Some probabilities are likely to show a considerable effect, while others will hardly have any influence. For the less influential probabilities, the initial assessments may suffice. For the more influential probabilities, on the other hand, refinement may be worthwhile; for example, more elaborate 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 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, we refer the reader to an extended paper [Coupé *at al.*, 2000].

5 Evaluation of the elicitation method

We used our newly designed method for probability elicitation from domain experts in the construction of the probabilistic part of the oesophagus network. In this section, we evaluate the use of our method. More specifically, we comment upon the observations made by the domain experts involved.

5.1 Using the method

In the first interview with our 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 the various assessments, and that, if necessary, we would try to refine the most influential ones later on. The basic idea of sensitivity analysis was explained in some detail 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 conditional probabilities required for the part of the oesophagus network outlined in Section 2 took five interviews of approximately two hours each over a period of fifteen months. Each interview focused on a small coherent part of the network. Prior to each interview, the elicitors spent some ten hours preparing the fragments of text and associated response scales to be presented to the experts; after the interview, it took the elicitors two to five hours to process the obtained assessments. The new method allowed the domain experts to give their assessments at a rate of 150 to 175 probabilities per hour; the remaining time was spent on explanation and instruction.

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 domain 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, rather 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 reactive growth of cancer cells, along with information recalled from literature, enabled them to provide the required assessments without much difficulty.

With respect to the response scale used for marking assessments, 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 conditional 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 for marking their assessment on the scale. 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 on the scale then helped them to determine the position that they felt expressed the probability 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 appropriate anchor or put a mark close to the anchor on the scale. In the later interviews, they considered the entire response scale, marked 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. Note that these observations pertain to the lower half of the scale only. We would like to add that our response scale hardly accommodates for indicating extreme probability assessments, that is, assessments very close to 0% or 100%. There are no anchors close to zero and one hundred percent probability on the scale since only very few subjects in our study had generated extreme verbal expressions. The domain experts never seemed to want to express such extreme assessments either. When asked about this, the experts 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 are 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 later interviews, the domain experts were able to verify the coherence of their assessments for the same conditional distribution 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* that models the depth of invasion of an oesophageal carcinoma 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 carcinoma has invaded into the oesophageal wall and the worse off the patient is. For the variable *Invasion*, several conditional probabilities were required, pertaining to different shapes and varying lengths of the carcinoma. Upon assessing these probabilities, the domain experts started with the probabilities for the depth of invasion of a *polypoid* oesophageal carcinoma with a length of less than 5 centimeters.

They subsequently indicated that patients with *ulcerating* tumours of this length were 10% worse off with regard to the depth of invasion of the carcinoma than patients with similar polypoid tumours. They thus explicitly related 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 correct way, with the trends provided by our domain experts. The method is best explained in terms of the example trend given above. Suppose that, given a *polypoid* oesophageal carcinoma of less than 5 centimeters 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, $i = 1, 2, 3, 4$, constitute the *anchor distribution* that is to be adjusted by the indicated trend to compute the probabilities for the related distribution. After consultation with our domain experts, we interpreted the specified trend as follows: *10% of the patients with a polypoid tumour of less than 5 centimeters with Ti for its depth of invasion would have had $Ti + 1$ for the depth of invasion if the tumour was an ulcerating tumour*, $i = 1, 2, 3$. The basic idea of the interpretation of the trend is depicted in Figure 6. For the probability assessments y_1, y_2,

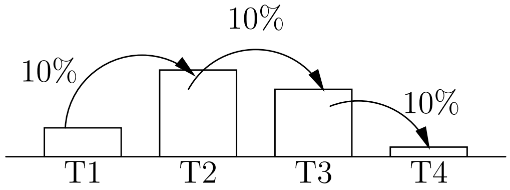

Figure 6: A schematic representation of handling trends.

y_3, and y_4 for the different values of the variable *Invasion* given an *ulcerating* oesophageal carcinoma of less than 5 centimeters, we find

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

It is readily verified that y_1, y_2, y_3, and y_4 lie between 0 and 1, and together sum up to 1. In addition, it will be evident that this method for 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 Evaluation of the elicited probabilities

To assess the quality of the probabilities obtained with our new elicitation method, we conducted an evaluation study of the oesophagus network. In the study, we used data from patients from the Antoni van Leeuwenhoekhuis diagnosed with oesophageal carcinoma. In Section 6.1, we analyse the probabilities obtained; we compare them with the data in Section 6.2. In Section 6.3, we study the 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 carcinoma from the network; we then compared the computed stage with 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, a total of 900 probabilities were required. The number of probabilities to be assessed per variable ranged from 3 to 144, constituting a total of 267 (conditional) probability distributions. Many of the assessments we obtained from our domain experts equaled 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, however, that 144 of these probabilities pertain to the deterministic variable that models a carcinoma's stage, that is, 35% of the zeroes and ones constitute the (degenerate) 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 discussed in Section 5.2. 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. Figure 7(a) summarises the frequencies of all assessments obtained, be it directly or indirectly; we have restricted the figure to the assessments not equal to zero or one. Figure 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. The two tables from Figure 8 reveal the ten most frequently specified assessments, counted over all assessments and over the direct assessments only.

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 from Figure 8 further reveal that among the ten most often specified assessments, there are four anchors from the response scale: 0, 0.15, 0.85, and 1.00. Among the ten most frequently specified direct assessments, there even are six anchors: 0, 0.15, 0.25, 0.75, 0.85, and 1.00. These findings are consistent with the often reported observation that

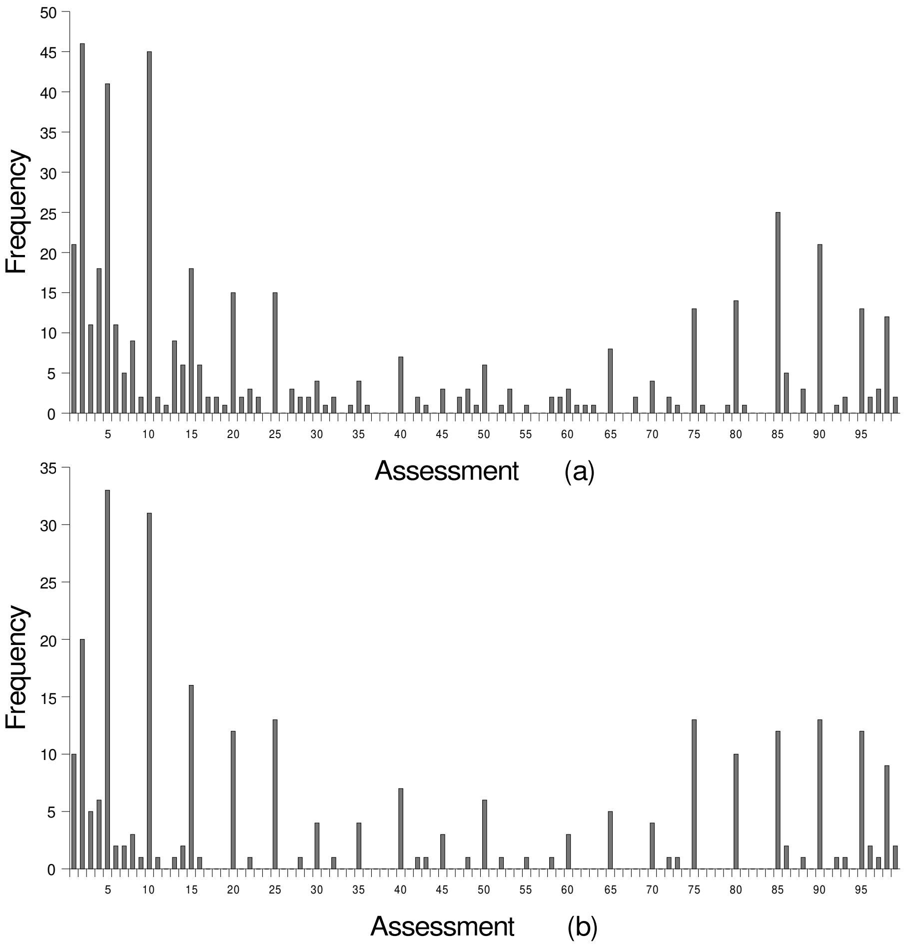

Figure 7: The distribution of all assessments obtained (a) and of the assessments that were specified directly (b), 0% and 100% excluded.

the external stimulus used, in our case the response scale, plays a dominant role in the elicitation process. 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 have given this assessment only seven times.

6.2 A comparison with 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 network. Our efforts to compose such a data collection, however, had resulted in data from historical records of 185 patients diagnosed with oesophageal cancer from the Antoni van Leeuwenhoekhuis. As these data had not been used for probability assessment, we could

assessment	*frequency*
0	312
1.00	100
0.02	46
0.10	45
0.05	41
0.85	25
0.01	21
0.90	21
0.04	18
0.15	18

(a)

assessment	*frequency*
0	272
1.00	92
0.05	33
0.10	31
0.02	20
0.15	16
0.25	13
0.75	13
0.90	13
0.85	12

(b)

Figure 8: The ten most frequent assessments (a) and the ten most frequent direct assessments (b).

now exploit them for evaluation purposes. In this section, we compare the probabilities given by our domain experts with estimates from these data. Before doing so, however, 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. Since the historical records dated back to between 1978 and 1985, and the experts did not scrutinise the data prior to assessing the required probabilities, we concluded that the data were independent enough to render the evaluation results 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 computed: 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 in the data collection. The data 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 matched the estimates that we obtained from the data, we computed a 95% confidence interval for each of the 368 probability estimates. The 95% confidence interval of a specific estimate is the interval in which the 'true' probability lies with 95% certainty; the length of the confidence interval thus quantifies the uncertainty in the estimate. For a probability estimate p, its 95% confidence interval is approximated as

$$\left(-1.96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}, +1.96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}\right)$$

where n is the number of patients whose data have been used in the computation of the estimate p. Note that the larger the number of patients on which the estimate is based, the smaller the estimate's 95% confidence interval. The confidence intervals that we thus obtained for our probability estimates were rather large as a result of data sparseness: the intervals had an average length of 0.25. For 250 of the 368 estimates, the 95% confidence interval included the assessment that we had elicited from the experts. So, from the assessments that could be compared with the data, 68% more or less matched the computed probability estimates.

As discussed 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, with probability estimates from the data. To determine the goodness of fit of a specific estimated distribution on the same distribution specified by the experts, we conducted a number of χ^2-tests. A χ^2-test builds upon a χ^2-distribution for the difference between the two probability distributions that are compared. This difference is measured as

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

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 can conclude that the two distributions do not match. Figure 9 summarises the match results that we obtained from the various χ^2-tests.

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

Figure 9: The number of matching anchor and indirectly computed distributions.

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 closely matched the data, we may conclude that the direction as well as the percentage of adjustment that were indicated by our domain experts are correct. 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 was correct. 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 9 trends, that is, for 39% of the trends specified by the domain experts, the indicated direction and percentage of adjustment are correct. Alternatively, 61% of the trends appear to be incorrect. Upon examining the fourteen apparently incorrect trends, we found that for four of them a trend seemed to be reflected in the data: 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, no conclusive statements with regard to the correctness of the specified trends can be made.

6.3 The quality of the network

To conclude our evaluation of the elicited probabilities, we conducted a study of the oesophagus network with data from 185 patients diagnosed with oesophageal carcinoma from the Antoni van Leeuwenhoekhuis. The study once again focused on the part of the network that provides for establishing the stage of a patient's carcinoma; the stage of an oesophageal carcinoma can be either I, IIA, IIB, III, IVA, or IVB, in the order of progressive disease. Unfortunately, for 29 patients from our data collection the stage of their carcinoma was not recorded, leaving us with 156 patients for evaluation.

In a first evaluation of the oesophagus 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 carcinoma from the network and compared it with the stage recorded in the data. Figure 10 shows the results from this first evaluation. For 80 of the 156 patients, the stage of the carcinoma 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 concluded 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 [Berner *et al.*, 1994].

Taking the results from the first evaluation as a point of departure, we carefully examined the data of the patients for whom the probabilistic network returned a stage different from the recorded one. We identified three major sources of mismatch which could

		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

Figure 10: The results from the first evaluation.

largely be attributed to problems with the data. 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 source of mismatch. For example, for some patients a deeper invasion of the carcinoma 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 into the network, a stage different from the recorded one was established. The third major source of mismatch was found in the way 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 these metastases, which 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 was very unlikely that this test had been performed in the majority of the patients from our data collection, we concluded that some variables modeling diagnostic tests were missing from our network.

Building upon the above observations, we decided to perform a second evaluation of the oesophagus network. For this purpose, we first extended the network with three extra statistical variables pertaining to diagnostic tests. In close consultation with our domain experts, we had identified two additional diagnostic tests for establishing the presence or absence of metastases in the lymph nodes near the truncus coeliacus and one for establishing the presence or absence of lymphatic metastases in the cervix. In addition, we corrected the erroneous stages in the data, that is, as far as they had been identified by our experts in the first evaluation of the network.

In the second evaluation of the oesophagus network, we entered for each patient the available symptoms and test results, as before. If no tests were explicitly specified for facts with regard to lymphatic metastases in the cervix or near the truncus coeliacus, we entered these facts as test results for the 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 evaluation are shown in Figure 11. Figure 12 summarises the results per stage.

		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

Figure 11: The results from the second evaluation.

Figure 12(a) shows, per stage from the data, the percentage of patients for whom the network computed the same stage; these percentages can be interpreted as the sensitivity per stage of our network to the patient data. Figure 12(b) shows, per stage computed from the network, the percentage of patients for whom the data records the same stage; these percentages constitute the predictive value per stage of the network's outcome. Figure 11 reveals that for 132 of the 156 patients, the stage of the carcinoma 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 established the correct stage for 85% of the patients.

stage from data	*matched by network*
I	100%
IIA	97%
IIB	0%
III	74%
IVA	90%
IVB	88%

(a)

stage from network	*matched by data*
I	67%
IIA	76%
IIB	–
III	76%
IVA	100%
IVB	100%

(b)

Figure 12: The results from the second evaluation, detailed per stage from the data (a) and per stage computed from the network (b).

7 Concluding observations

A decision-support system is being developed for patient-specific therapy selection for oesophageal carcinoma. The kernel of the system is a probabilistic network describing the characteristics of oesophageal carcinoma and the pathophysiological processes of invasion and metastasis. In the development of our network, we found that probability elicitation can be a major obstacle. Building upon our negative experiences with existing methods, we designed a new method for eliciting probabilities from domain experts. 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 method for eliciting the probabilities required for the oesophagus network and evaluated its use with the domain experts involved. The 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.

Using data from 185 patients, we evaluated the oesophagus network. A first evaluation revealed various sources of mismatch between the stage of a patient's carcinoma as recorded in the data and the one computed from the network. To a large extent, the mismatches could be attributed to anomalies in the data. We feel that this is not uncommon in evaluation studies like the present one. Additionally, the first evaluation served to identify a small number of variables missing from the network. After correcting the anomalies in the data and including the missing variables, 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 the study are quite encouraging. We are currently performing a sensitivity analysis of the network to identify the most influential assessments. Also, we are investigating the full network's ability to predict the outcome of treatment. We hope to report on our results in the near future.

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 little 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 with the method makes this time certainly well spent.

References

[Berner *et al.*, 1994] E.S. Berner, G.D. Webster, A.A. Shugerman, J.R. Jackson, J. Algina, A.L. Baker, E.V. Ball, C.G. Cobbs, V.W. Dennis, E.P. Frenkel, L.D. Hudson, E.L.

Mancall, C.E. Rackley, and O.D. Taunton (1994). Performance of four computer-based diagnostic systems. *The New England Journal of Medicine*, vol. 330, pp. 1792 – 1796.

[Brun & Teigen, 1988] W. Brun and K.H. Teigen (1988). Verbal probabilities: ambiguous, context-dependent, or both ? *Organizational Behavior and Human Decision Processes*, vol. 41, pp. 390 – 404.

[Budescu *et al.*, 1988] D.V. Budescu, S. Weinberg, and T.S. Wallsten (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, vol. 14, pp. 281 – 294.

[Coupé *et al.*, 2000] V.M.H. Coupé, L.C. van der Gaag, and J.D.F. Habbema (2000). Sensitivity analysis: an aid for belief-network quantification. *Knowledge Engineering Review*, vol. 15, pp. 1 – 18.

[Druzdzel & Van der Gaag, 1995] M.J. Druzdzel and L.C. van der Gaag (1995). Elicitation of probabilities for belief networks: combining qualitative and quantitative information. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, CA, pp. 141 – 148.

[Druzdzel & Van der Gaag, 2000] M.J. Druzdzel and L.C. van der Gaag (2000). Building probabilistic networks: Where do the numbers come from ? *IEEE Transactions on Knowledge and Data Engineering*, to appear.

[Gigerenzer & Hoffrage, 1995] G. Gigerenzer and U. Hoffrage (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, vol. 102, pp. 684 – 704.

[Jensen, 1995] A.L. Jensen (1995). Quantification experience of a DSS for mildew management in winter wheat. In: M.J. Druzdzel, L.C. van der Gaag, M. Henrion, and F.V. Jensen. *Working Notes of the Workshop on Building Probabilistic Networks: Where Do the Numbers Come From ?*, pp. 23 –31.

[Jensen, 1996] F.V. Jensen (1996). *An Introduction to Bayesian Networks*, UCL Press, London.

[Kahneman *et al.*, 1982] D. Kahneman, P. Slovic, and A. Tversky (1982). *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

[Merz *at al.*, 1991] J.F. Merz, M.J. Druzdzel, and D.J. Mazur (1991). Verbal expressions of probability in informed consent litigation. *Medical Decision Making*, vol. 11, pp. 273 – 281.

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

[Renooij & Witteman, 1999] S. Renooij and C.L.M. Witteman (1999). Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, vol. 22, pp. 169 – 194.

[Von Neumann & Morgenstern, 1953] J. von Neumann and D. Morgenstern (1953). *The Theory of Games and Economic Behavior*, Wiley, New York, 3rd edition.

[Von Winterfeldt & Edwards, 1986] D. von Winterfeldt and W. Edwards (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.

[Wallsten *et al.*, 1993] T.S. Wallsten, D.V. Budescu, and R. Zwick (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, vol. 39, pp. 176 – 190.