The Effects of Disregarding Test Characteristics in Probabilistic Networks

Linda C. van der Gaag, C.L.M. Witteman, Silja Renooij, and M. Egmont-Petersen

Institute of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
{linda,cilia,silja,michael}@cs.uu.nl

Abstract. In most medical disciplines, the results from diagnostic tests are not unequivocal. To capture the uncertainties in test results, the notions of sensitivity and specificity of diagnostic tests have been introduced. Although the importance of taking these test characteristics into account in medical reasoning is stressed throughout the literature, they are often not modelled explicitly in real-life probabilistic networks. In this paper, we study the effects that disregarding the characteristics of diagnostic tests can have on the performance of a probabilistic network. We feel that the effects that we observed in a real-life network for the staging of oesophageal cancer, are likely to be found in networks for other applications in medicine as well.

1 Introduction

Since their introduction, probabilistic networks have become increasingly popular in medical decision support. A probabilistic network is a statistical model comprised of a graphical structure and an associated set of probabilities [1]. The graphical structure models the statistical variables that are relevant in the field of application, along with the influential relationships between them; the strengths of the relationships are captured by conditional probabilities. Probabilistic networks have a number of methodological advantages compared to other types of statistical model used in medical reasoning. Probabilistic networks can for example adequately deal with data from patients in whom only a subset of the relevant variables have been observed.

Medical reasoning with a probabilistic network amounts to entering a patient's symptoms and test results into the network. The network then computes the most likely diagnosis for the patient, or another outcome of interest. In most medical disciplines, a patient's symptoms and test results are not unequivocal. As an example, we consider an X-ray of a patient's chest to establish the presence or absence of lung cancer. A physician interpreting the X-ray may easily overlook a small tumour and falsely state a negative result. On the other hand, the physician may state a false positive result based on a phantom image. The uncertainties in a test's results are captured by the *sensitivity* and *specificity characteristics* of the test. The sensitivity of a test is defined as the probability

S. Quaglini, P. Barahona, and S. Andreassen (Eds.): AIME 2001, LNAI 2101, pp. 188–198, 2001.

that a positive test result is found in a patient who actually has the disease. The test's specificity is the probability that the test yields a negative result for a patient without the disease. The notions of sensitivity and specificity are well-known in the field of medical decision making; textbooks in fact stress the importance of taking test characteristics into account in medical reasoning in order to avoid misdiagnosis [2,3].

Probabilistic networks allow for modelling test characteristics by distinguishing between variables that represent test results on the one hand and variables that model the (generally unobservable) truth on the other hand, and thus provide the means for dealing with the uncertainties in test results. In many real-life probabilistic networks, however, the sensitivity and specificity characteristics of tests are not explicitly modelled, not even in medical disciplines where these characteristics are readily available. Building a probabilistic network is a hard and time-consuming task and, in our experience, incorporating test characteristics contributes significantly to its difficulty. The main obstacle is not so much in capturing the characteristics themselves but rather in obtaining the probabilities for the variables that model the true values. We feel that this observation explains the absence of test characteristics from many real-life networks.

In this paper, we investigate the effects that disregarding the sensitivity and specificity characteristics of diagnostic tests can have on the performance of a probabilistic network. To this end, we study a real-life network for the staging of oesophageal cancer. We compare the performance of the complete network that captures full test characteristics, with the performance of a reduced network from which these characteristics have been removed. We feel that the various effects that we observe in the performance of these networks are likely to be found for probabilistic networks for other applications in medicine as well.

In Sect. 2, we briefly introduce our probabilistic network for the staging of oesophageal cancer. In Sect. 3, we present the design of our study. In Sect. 4, we describe the results from evaluations of the complete network and of the reduced network, using real-life patient data; these results are discussed in depth in Sect. 5. The paper ends with our concluding observations in Sect. 6.

2 The Oesophagus Network

The *oesophagus network* for the staging of oesophageal cancer has been constructed with the help of two experts in gastrointestinal oncology from the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis [4]. The network describes the presentation characteristics of an oesophageal tumour, such as its location in the oesophagus, and its histological type, length, and macroscopic shape. In addition, the network describes the pathophysiological process underlying the tumour's invasion into the oesophageal wall and adjacent organs. It further describes the process of metastasis. The depth of invasion and the extent of metastasis are summarised in the tumour's *stage*; this stage can be either I, IIA, IIB, III, IVA, or IVB, in the order of advanced disease. The oesophagus network includes 42 variables, for which almost 1000 probabilities have been

specified by our experts. The network is depicted in Fig. 1, which also shows the prior probabilities computed per variable.

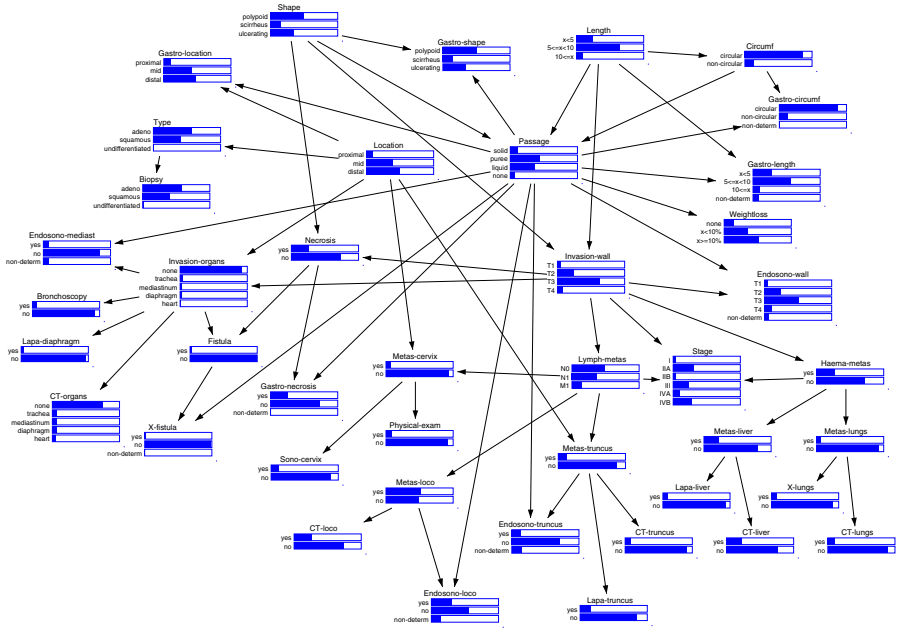
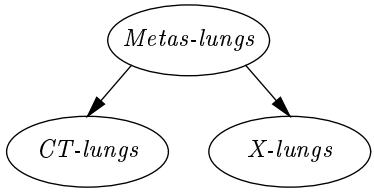


Fig. 1. The oesophagus network.

To establish the stage of a patient's oesophageal tumour, typically a number of diagnostic tests are performed, ranging from multiple biopsies of the tumour to a CT-scan of the patient's chest. The network includes 23 variables to model test results. The sensitivity and specificity characteristics of the tests are captured by the associated conditional probabilities. As an example, Fig. 2(b) shows the probabilities of the results of an X-ray of a patient's chest, given the actual presence or absence of lung metastases: the X-ray is stated to have a sensitivity of 0.85 and a specificity of 0.98. The diagnostic tests included in the oesophagus network differ considerably with respect to their sensitivity and specificity characteristics, that is, not every test result obtained for a patient is equally reliable.

3 The Study

To study the effects of disregarding the sensitivity and specificity characteristics of diagnostic tests, we compare the performance of the oesophagus network described in the previous section with that of a reduced network from which the 23 variables representing test results have been removed. We refer to the former as



(a)

		Metas-lungs	
		yes	no
X-lungs	yes	0.85	0.02
	no	0.15	0.98

(b)

Fig. 2. A fragment of the oesophagus network (a) and some associated probabilities (b).

the *complete network* and to the latter as the *reduced network*. The performance of the two networks is compared using real-life patient data. For each patient, the most likely stage, given the available symptoms and test results, is computed from the two networks; the computed stages are then compared with the stage recorded in the data.

In the complete oesophagus network, test results are entered as values for the appropriate variables. For example, if a gastroscopic examination of a patient's oesophagus reveals a tumour length of less than 5 centimeters, this result is entered for the variable *Gastro-length*. From the reduced network, however, the variables modelling test results have been removed. As a consequence, no distinction can be made between a test result and the true value. Test results are therefore entered as if they were true values. For example, the result mentioned above is entered for the variable *Length* which models the true tumour length.

For variables for which a *single* test is performed, such as the variable *Length*, test results can be entered unambiguously into the reduced network. For several variables, however, two or more tests are available. For example, to establish the presence or absence of metastases in a patient's lungs, an X-ray or a CT-scan can be made, as indicated in Fig. 2(a). The results from such multiple tests can be conflicting, that is, it is possible that different results are yielded by the CT-scan and the X-ray. In case of such conflicting test results, we have to decide which result to enter. A range of different strategies can be designed for this purpose:

- a result is picked randomly (the *random-test strategy*);
- a result from a more reliable test is preferred over a less reliable result (the *reliable-test strategy*);
- a positive result is preferred over a negative one (the *positive-test strategy*).

The *reliable-test strategy* is closest to the reasoning strategy exhibited by the complete network: it takes the sensitivity and specificity of the various tests into account, even though these characteristics are not explicitly modelled. The *positive-test strategy*, on the other hand, seems to be the closest to the reasoning behaviour of the gastroenterologists who helped in the network's construction and collected the data. In our study, we therefore focus on these two strategies.

We would like to note that, in case of conflicting test results, the complete oesophagus network takes *all* results into account: it weighs the uncertainties involved and combines the results into an overall uncertain value. The network is said to exhibit *compensatory* reasoning behaviour. By removing the variables that represent test results, we have shifted part of this compensatory behaviour to outside the network. Our strategies for entering test results are *non-compensatory* as they do not weigh and combine results. With these strategies, the reduced network exhibits a reasoning behaviour that is *non-compensatory* to at least some extent.

In our study, we use data from the medical records of 156 patients from the Antoni van Leeuwenhoekhuis in the Netherlands diagnosed with oesophageal cancer. For these patients an average number of 12.9 test results are available, ranging from 6 to 19 results per patient.

4 The Results

The results from our study are summarised in Fig. 3. The matrix shown in Fig. 3(a) describes the results from the evaluation of the complete oesophagus network. The numbers on the diagonal of the matrix are the numbers of patients per stage, for whom the network yields the same stage as the one recorded in the data. Because, unfortunately, a *gold standard* for the staging of oesophageal cancer is wanting, we take the stages entered into the patients' medical records for our 'silver' standard of validity. The matrix then reveals that the network computes the correct stage for 110 patients, yielding a percentage correct of 71%, with a 95% confidence interval of (63.4%,77.7%).

The matrices shown in Fig. 3(b) and 3(c) describe the results from evaluating the reduced network, with the reliable-test strategy and the positive-test strategy, respectively. In contrast with the complete oesophagus network, the reduced network is not able to establish a stage for every patient from our data collection: for some patients, entering their data results in an inconsistency and an *error* is generated. The matrix from Fig. 3(b) reveals that the reduced network with the reliable-test strategy computes the correct stage for 66% of the patients, with a 95% confidence interval of (58.6%,73.5%); the data of 22 patients, or 14%, cannot be processed. With the positive-test strategy, the reduced network computes the correct stage for 76% of the patients, with a 95% confidence interval of (68.9%,82.4%); the data of 8 patients, or 5%, result in an error.

Fig. 3(d), to conclude, describes the difference in performance of the reduced network with the reliable-test strategy and with the positive-test strategy, respectively, for the patients for whom neither of the strategies results in an error. As all discrepancies occur below the diagonal of the matrix, it is readily seen that the positive-test strategy tends to result in higher stages being concluded than the reliable-test strategy.

		complete network						
		I	IIA	IIB	III	IVA	IVB	*error*
	I	**2**	0	0	0	0	0	0
	IIA	0	**37**	0	1	0	0	0
data	IIB	0	2	**0**	2	0	0	0
	III	1	17	0	**28**	0	1	0
	IVA	1	1	0	9	**28**	0	0
	IVB	0	1	0	6	4	**15**	0

(a)

		network, reliable test						
		I	IIA	IIB	III	IVA	IVB	*error*
	I	**1**	0	0	0	0	1	0
	IIA	0	**38**	0	0	0	0	0
data	IIB	0	2	**0**	2	0	0	0
	III	0	11	0	**29**	0	1	6
	IVA	0	2	0	4	**18**	0	15
	IVB	0	2	0	2	4	**17**	1

(b)

		network, positive test						
		I	IIA	IIB	III	IVA	IVB	*error*
	I	**1**	0	0	0	0	1	0
	IIA	0	**38**	0	0	0	0	0
data	IIB	0	2	**0**	2	0	0	0
	III	0	9	0	**33**	1	2	2
	IVA	0	1	0	4	**29**	0	5
	IVB	0	0	0	4	4	**17**	1

(c)

		network, reliable test					
		I	IIA	IIB	III	IVA	IVB
	I	**1**	0	0	0	0	0
	IIA	0	**50**	0	0	0	0
network,	IIB	0	0	**0**	0	0	0
positive	III	0	3	0	**36**	0	0
test	IVA	0	2	0	0	**22**	0
	IVB	0	0	0	1	0	**19**

(d)

Fig. 3. The performance of the complete network (a), of the reduced network with the *reliable-test* (b) and *positive-test strategies* (c), and a comparison of the latter two (d).

5 The Effects

In comparing the performance of the complete oesophagus network with that of the reduced network using the reliable-test and positive-test strategies, we observe several effects that can be attributed to test characteristics. These effects are discussed in Sect. 5.1 through 5.3; Sect. 5.4 addresses the overall effect.

5.1 The Effect of the Graphical Structure

The graphical structure of a probabilistic network portrays the influential relationships between its variables. More formally, the structure captures probabilistic dependences and independences [1]. Two variables A and B are *independent* given a set of observations if, on every trail between A and B, there is a variable Y for which one of the following conditions holds:

- Y is observed and on the trail there is at least one arc emanating from Y;
- Y has two incoming arcs on the trail and neither Y nor any of its descendants in the graphical structure is observed.

As the set of observed variables changes, so will the set of dependences and independences that are read from a network's graphical structure. For example, in Fig. 1, the variables *Shape* and *Length* are independent, yet become dependent

if a value for *Passage* is entered. On the other hand, the variables *Gastro-length* and *Stage* are dependent, indicating that entering the test result of a gastroscopic examination of a patient's oesophagus with respect to tumour length, will affect the probabilities of the various stages. If the value of *Invasion-wall* would be known, however, entering the test result would no longer have this effect, because *Gastro-length* and *Stage* would then have become independent.

Since a patient's test results are entered for different variables in the complete oesophagus network and in the reduced network, different dependences and independences are taken into account in the networks' computations. To illustrate this observation, we consider the test results and symptoms of a specific patient:

Biopsy = adeno *Gastro-circumf* = circular *Gastro-length* = $5 \leq x \leq 10$
Passage = puree *Gastro-shape* = polypoid *Gastro-location* = distal
Gastro-necrosis = no

When these data are entered into the complete network, each test result and symptom affects the probabilities of the various possible stages for the patient's tumour: the variable *Stage* is dependent of every variable mentioned. Furthermore, four extra *dependences* arise that cause additional interaction effects of the test results and symptoms on the probabilities for the variable *Stage*. For the reduced network, the patient's data are interpreted as

Type = adeno *Circumf* = circular *Length* = $5 \leq x \leq 10$
Passage = puree *Shape* = polypoid *Location* = distal
Necrosis = no

When these values are entered, some new *independences* arise: the variable *Stage* has become independent of the variables *Type* and *Circumf*. As a consequence, the results from the biopsies of the tumour and from the gastroscopic examination of the tumour's circumference are no longer taken into account in the computation of the most likely stage. Furthermore, compared with the complete network, only two of the additional dependences arise, as a result of which fewer interaction effects of the data are captured.

In the complete oesophagus network, a maximum of 23 test results and two symptoms can be entered. Regardless of which test results and symptoms are available, the probabilities for the variable *Stage* depend on each of them; the only exception occurs when a value for the variable *Passage* renders the variable *Stage* independent of *Weightloss*. In the reduced network, a maximum of 15 values can be entered. Entering these values may give rise to new independences, causing a number of test results and symptoms to be disregarded in the computations. For our data collection, between 25% and 45% of the test results and symptoms per patient are thus disregarded.

5.2 The Effect of the Non-compensatory Strategies

The complete oesophagus network takes the uncertainties in a patient's test results into account in its reasoning behaviour. The network exhibits compensatory behaviour in the sense that it carefully weighs the uncertainties involved. For the reduced network, we have shifted part of this compensatory behaviour

to outside the network. Since our strategies for entering test results are basically non-compensatory, the reduced network exhibits a reasoning strategy that is non-compensatory to at least some extent.

The way that people, including expert physicians, deal with uncertainties is studied in the field of cognitive science. Various studies have revealed that people interpret probabilistic information differently than probability theory prescribes. People, in fact, exhibit non-compensatory reasoning behaviour. As an example, we briefly review a study by D.M. Eddy [5], pertaining to the diagnosis of breast cancer. In the study, physicians were presented with the following information: the sensitivity of a mammography equals 0.90 and its specificity is 0.80. They were also told that one out of every 10 patients with similar complaints as a particular patient, indeed had breast cancer. The physicians were then asked to assess the probability that this patient, given a positive test result, had cancer. Many physicians in the study judged the probability to be 0.90; most others gave a probability over 0.50. However, when the requested probability is correctly calculated from the information presented, the result is 0.33, a much lower probability. The physicians who gave high probabilities apparently confused the conditional probability expressing sensitivity with its inverse: they thought that the probability of cancer given a positive test result must be equal to the probability of a positive test result given cancer. This confusion of the sensitivity of a test for the predictive value of a positive test result is often observed in cognitive studies. In practice, physicians tend to assume that diagnostic tests are quite sensitive and specific, and interpret a result as an unequivocal *yes* or *no*. They in fact prefer to act as if everything were deterministic [6].

Interpreting test results as true values can cause highly unlikely combinations of results to become truly conflicting. To illustrate this observation, we consider the test results that are mentioned in a particular patient's medical record. The record states *CT-organs* = none, indicating that on the CT-scan of the patient's chest no invasion of the tumour into adjacent organs is visible. For the patient is also stated *X-fistula* = yes, indicating that the X-ray shows an open connection between the oesophagus and the trachea. When these test results are entered into the complete oesophagus network, it weighs the uncertainties involved and combines them into an uncertain assessment for the variable *Invasion-organs*, which models the true invasion of the tumour beyond the oesophageal wall. In the reduced network, the test results are entered as *Invasion-organs* = none and *Fistula* = yes, regardless of the strategy used. Since a fistula can occur only if the oesophageal tumour has invaded the trachea, the two values constitute an inconsistency and an error is generated. While the complete network is able to handle highly unlikely combinations of test results by its compensatory reasoning behaviour, the reduced network, with its non-compensatory strategies, is not.

5.3 The Effect of Preferring Positive Test Results

Physicians are trained to save lives and, as a consequence, are *loss aversive* [7], that is, they will generally try to avoid losing a patient's life. Moreover, physicians have been found to look upon an omission, such as abstaining from surgery, which

results in a serious condition, as less bad than performing an intervention which leads to the same condition. In the field of oesophageal cancer, for example, we have observed our gastroenterologists, when assessing the stage of a patient's tumour, to go by a positive result, even if another related test was yet available or had yielded a negative result: they seem to prefer to base their decisions on the positive result in order to be 'better safe than sorry'. This observation coincides with the *confirmation bias* that is often observed in cognitive studies [8]. Positive information tends to get more attention than negative information; in fact, negative information is simply neglected when positive information is available. We feel that these biases arise from decision considerations and should not be taken into account in probabilistic reasoning.

The positive-test strategy used with the reduced oesophagus network complies with the confirmation bias described above and closely resembles our expert gastroenterologists' interpretation of patient data. The results from Sect. 4 indicate that the reduced network with this strategy computes the correct stage for a larger number of patients than the complete network and than the reduced network with the reliable-test strategy. From this observation, we cannot simply conclude that the reduced network with the positive-test strategy performs best. We recall that in the absence of a gold standard for the staging of oesophageal cancer, we have taken the stages from the patients' medical records for our silver standard of validity. This standard, although the best available, is known to be imperfect: the data entered by the gastroenterologists reflect their biases. Since we are comparing against an admittedly imperfect standard, one network may perform better than another while in fact it could be worse. This phenomenon is well-documented in the field of medical decision making [2]. Based upon this observation, we feel that from the evaluation results described in Sect. 4, we can only conclude that the reduced network with the positive-test strategy best fits the data.

5.4 The Overall Effect

The results from our study, as presented in Sect. 4, show that the reduced network using the reliable-test strategy computes the correct stage for 66% of the patients from our data collection; the complete network yields a percentage correct of 71% and the reduced network with the positive-test strategy results in 76% of the patients being correctly classified. Although only moderately significant, we would like to comment on these differences. The difference in performance between the complete network and the reduced network with the reliable-test strategy can be explained from the effect of the graphical structure and from the effect of using a non-compensatory strategy for entering test results. We feel that these effects basically constitute the overall effect of disregarding test characteristics. These effects come into play also for the reduced network with the positive-test strategy. The overall effect of disregarding the characteristics of diagnostic tests, however, is now dominated by the effect of the strategy used because it closely matches the biases in the data.

6 Conclusions

We have investigated the effects that disregarding the sensitivity and specificity characteristics of diagnostic tests can have on the performance of a probabilistic network. To this end, we have compared the performance of a *complete network* for the staging of oesophageal cancer including full test characteristics, with that of a *reduced network* from which these characteristics have been removed. In the reduced network, test results are entered as if they were true values. For entering conflicting test results, we have introduced the *reliable-test* and *positive-test strategies*.

In our study of the oesophagus network, we have observed various effects of disregarding test characteristics that are likely to be found in most applications of probabilistic networks in medicine. We have found that the strategy for entering test results highly influences the reduced network's performance. Also, while the reliable-test strategy should perform better because it more closely resembles correct probabilistic reasoning, it is the positive-test strategy that gives the better result. We have attributed this paradoxical observation to the use of an imperfect standard of validity. We have further found that entering test results as true values results in a large number of errors for real-life patient data. Since a probabilistic network should be able to handle the data of real patients, such errors are highly unwished-for. If many errors are generated, the value of a network for clinical practice rapidly decreases. Based upon our observations, we believe that for most applications in medicine the characteristics of diagnostic tests should be explicitly modelled in a probabilistic network.

Acknowledgements. This research has been (partly) supported by the Netherlands Computer Science Research Foundation with financial support from the Netherlands Organisation for Scientific Research (NWO). We are most grateful to Babs Taal and Berthe Aleman from the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, who spent much time and effort in the construction of the oesophagus network.

References

1. F.V. Jensen (1996). *An Introduction to Bayesian Networks*. London: UCL Press.
2. R.H. Fletcher, S.W. Fletcher, E.H. Wagner (1996). *Clinical Epidemiology. The Essentials*, 3rd ed. Baltimore: Williams & Wilkins.
3. D. von Winterfeldt, W. Edwards (1986). *Decision Analysis and Behavioral Research*. New York: Cambridge University Press.
4. L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, B.G. Taal (2001). Probabilities for a probabilistic network: A case-study in oesophageal carcinoma, submitted for publication.
5. D.M. Eddy (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In: D. Kahneman, P. Slovic, A. Tversky (editors). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

6. J. Baron (1994). *Thinking and Deciding*, 2nd ed. Cambridge, UK: Cambridge University Press.
7. D. Kahneman, A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, vol. 47, pp. 263 – 291.
8. J.St.B.T. Evans, D.E. Over (1996). *Rationality and Reasoning*. Hove, UK: Psychology Press.