Building Bayesian Networks through Ontologies

Eveline M. Helsper[1] and **Linda C. van der Gaag**[1]

Abstract. To support building and maintaining knowledge-based systems for real-life application domains, sophisticated knowledge-engineering methodologies are available. As more and more Bayesian networks are being developed for complex applications, their construction and maintenance calls for the use of tailor-made knowledge-engineering methodologies. We have designed such a methodology and have studied its use within the domain of oesophageal cancer. Based upon expert knowledge and a previously constructed Bayesian network, we have built an ontology for this domain, from which we have constructed, in a sequence of steps, a new network. The use of our methodology has allowed us to address, in a structured fashion, the various intricate modelling issues involved.

1 INTRODUCTION

Building and maintaining a knowledge-based system for a complex real-life application domain is a hard and time-consuming process. Knowledge has to be elicited from domain experts and carefully captured in the representation formalism used by the system. Nowadays, sophisticated knowledge-engineering methodologies are available to support this process [1].

As demonstrated by an increasing number of applications in a range of domains, more and more knowledge-based systems use the formalism of Bayesian networks for their knowledge representation. A Bayesian network consists of a graphical structure, encoding statistical variables from the domain of application along with the influential relationships between them, and an associated numerical part, encoding a joint probability distribution over these variables [2]. Building such a network involves three basic tasks [3]. First, the statistical variables that are important in the domain must be identified, along with their possible values. Secondly, the relations between these variables must be identified and expressed in a graphical structure. The last task is to obtain the probabilities that are required for the network's numerical part. These tasks are typically performed with the help of domain experts.

Although building network-based systems resembles building knowledge-based systems in general, constructing a Bayesian network involves specific modelling issues, such as modelling domain concepts as statistical variables, that require more tailored methodologies. While the size and complexity of real-life networks have increased over the last decade, however, the literature on engineering network-based systems has not kept pace. Recently, an overall systems-engineering methodology has been advocated [4, 5]. We feel that such a methodology should indeed be adopted, but that it should be further detailed to arrive at a tailor-made methodology.

In this paper, we propose a knowledge-engineering methodology for building and maintaining Bayesian networks. Our methodology falls into line with more general methodologies that build upon the use of a knowledge model. We propose to model the knowledge of an application domain into an ontology and derive from this ontology, in a sequence of steps addressing modelling issues, the graphical structure of a Bayesian network. For quantifying the structure, we propose the use of currently available probability elicitation methods.

We have studied the use of our methodology within the domain of oesophageal cancer. We have constructed an ontology for this domain, based upon expert knowledge and upon a real-life Bayesian network that we had constructed before without the use of any specific methodology. We have derived various alternative graphical network structures from this ontology. The use of our methodology has allowed us to address, in a structured fashion, the intricate modelling issues involved. Comparison of the new graphical structures with the structure of the original network has in fact uncovered some awkward modelling decisions in the original network.

In this paper we introduce our knowledge-engineering methodology tailored to building Bayesian networks. As our methodology has not yet been fully developed and validated, we focus our presentation on its use within our application domain. In Section 2, we briefly introduce the domain of oesophageal cancer. In Section 3, we outline the basic idea of our methodology. In Section 4, we describe the ontology that we have constructed for our application domain. In Section 5, we show how alternative graphical structures can be derived from this ontology. The paper ends with our concluding observations.

2 THE OESOPHAGUS NETWORK

As a consequence of a lesion of the oesophageal wall, a tumour may develop in a patient's oesophagus. The various characteristics of the tumour, including its length and shape, influence its growth. The tumour typically *invades* the oesophageal wall and, upon further growth, may affect such neighbouring organs as the trachea. In time, the tumour may result in secondary tumours, or metastases, in lymph nodes and in such other organs as the liver and the lungs. A distinction is made between *lymphatic metastases* and *haematogenous metastases* that result from conveyance of cancer cells via the lymph vessels and via the blood vessels, respectively. The depth of invasion and the extent of the metastases, which are summarised in the cancer's *stage*, are indicative of the effects and complications to be expected from the different available therapeutic alternatives.

With the help of two experts in gastrointestinal oncology, we have built a Bayesian network for the staging of a patient's cancer of the oesophagus [6]. The network includes a graphical structure encoding statistical variables and the probabilistic relationships between them. The variables represent the concepts that are relevant for establishing the stage of a patient's cancer. The probabilistic influences between

[1] Institute of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands. Email: {eveline,linda}@cs.uu.nl

the variables are represented by directed links, or arcs. The set of arcs, more formally, captures probabilistic independence. Two variables are said to be independent given available observations if every chain between the two variables contains an observed variable with at least one emanating arc, or a variable with two incoming arcs such that neither the variable itself nor any of its descendants in the graph have been observed. The strengths of the represented influences are indicated by conditional probabilities. The oesophagus network includes some 40 statistical variables and a thousand probabilities.

3 AN OVERVIEW OF THE METHODOLOGY

Most currently available knowledge-engineering methodologies, roughly speaking, propose to develop a *knowledge model* in which domain knowledge is captured, and to use this model for constructing a *knowledge base*. Capturing knowledge directly in a knowledge base may result in a representation from which the domain knowledge is not easily recognizable as a result of the modelling decisions taken. First developing a separate knowledge model may thus prevent discrepancies that would seriously hamper the system's construction and maintenance. In our methodology we have adopted the use of a knowledge model. Because a Bayesian network cannot represent procedural knowledge, we have focused on the declarative domain knowledge only; providing for the procedural knowledge to be incorporated in a network-based system is subject to further research. As in deriving a network from a knowledge model numerous issues play an important role that are specific for Bayesian networks, we focus in this paper explicitly on the exploitation of the model.

In constructing a knowledge model, domain knowledge is acquired, usually from experts, and explicitly captured. For specifying the knowledge, which typically consists of concepts and the relations between them, we adopt the use of ontologies in our methodology [7]. Our ontology contains all knowledge that is to be captured in a network, as well as the available meta-level and background knowledge. It represents the knowledge in a structured fashion, for example in depictions, understandable for both the knowledge engineers and experts involved. It thus documents elicited domain knowledge and can be used as a basis for communication during engineering.

From the ontology that results from the previous phase of our methodology, a graphical structure is derived that after quantification should constitute a network that faithfully captures the relevant domain knowledge. This derivation takes place in two phases. First, the knowledge that is directly relevant for the network is selected from the ontology; the remainder of the ontology serves as background knowledge. The central concepts and relations from the selected knowledge are then combined into a single depiction. Based upon this depiction, a graphical structure is derived that adheres to the syntax of Bayesian networks. To this end, the domain concepts from the depiction are translated into statistical variables, that is, into variables with an exhaustive state space of mutually exclusive, discrete values; the relations from the depiction are subsequently translated into arcs between variables. In this phase, the translation of relations is performed without paying much attention to the fact that the arcs should correctly capture probabilistic independence.

Secondly, the graphical structure resulting from the previous phase is improved and optimised. As obtaining the probabilities is generally the bottleneck in developing a Bayesian network, the structure is first restricted to include only variables for which probabilities can be reasonably obtained, either from data or from experts. The arcs in the structure are then investigated as to whether or not they correctly capture the probabilistic independences that hold in the application domain. Arcs may have to be added or reversed. The resulting structure now faithfully represents the domain knowledge, but may be suboptimal from a practical point of view. To guarantee a feasible running time of probabilistic inference, the number of incoming arcs per variable may have to be reduced, for example, by removing weak dependences or by divorcing parents [2, 5]. Furthermore, the state spaces of some of the variables may have to be restricted. Note that both types of optimisation also help in reducing the number of probabilities that have to be assessed for the graphical structure. To fine-tune the structure, the optimisation steps are iterated.

In the derivation of a graphical structure from the ontology, decisions must be taken on numerous issues. Different decisions can result in different alternative structures. Since building a Bayesian network is a creative process, there is not always good reason to prefer one alternative to another. Especially in the optimisation phase, there is a trade-off between the desire for a rich network that captures the intricacies of the domain and the running time of probabilistic inference. The outcome of the trade-off will then depend on the requirements of the application. As not all trade-offs can be resolved as they arise, we propose to maintain several alternative structures, that are pruned as they are further developed. An issue to take into account upon pruning is that preserving the structure and contents of the model, or ontology, as much as possible during the derivation contributes to the maintainability of the resulting system [1].

We would like to note that the depiction of the selected knowledge as well as the alternative graphical structures can be quite large. We therefore propose to address small, semantically meaningful units of knowledge at a time. The consequences of any decision for a single unit on the other units should then be carefully taken into account throughout all the phases of the network's construction.

The quantification of a graphical structure is not yet fully supported by our methodology. We feel that a domain's uncertainties, and hence a network's quantification, should be taken into account already when building the ontology. This is subject to further research, however. For now we would like to mention that feasible methods are currently available that can be applied for quantifying the graphical structure that results from the earlier phases of our methodology [8].

4 THE OESOPHAGUS ONTOLOGY

In our methodology, the knowledge model used is an *ontology*. The term ontology refers to an explicit specification of the domain knowledge that is shared, for example by the experts and knowledge engineers involved in a system's construction [9]. An ontology specifies the knowledge that is explicitly captured in the system as well as the more implicit background knowledge of the domain and the meta-level knowledge of its regularities. Our main goal in developing an ontology is to make all elicited domain knowledge explicit. We have constructed an ontology for the field of oesophageal cancer. We would like to note that, although our ontology is based on the oesophagus network and the knowledge that has been elicited for the network's construction, it has not yet been validated in detail.

In developing an ontology, the representation language to be used should be chosen with care. It should introduce as little bias as possible in the contents and structure of the captured knowledge; otherwise, the ontology may not properly reflect the intricacies of the domain and may introduce errors in the future system. The language of Bayesian networks fails to meet this criterion, as important information may be lost in the translation of domain concepts and relations. We feel moreover, that an ontology that is expressed as a Bayesian network cannot serve as a communication medium between domain experts and knowledge engineers, mainly because the conceptual distance between the ontology and the way the experts think and talk about their domain would be too large. For the oesophagus ontology, we have chosen a semi-informal representation language that includes tables, graphs, and natural language. We feel that a more formal representation language would be less suitable, since using formal languages is uncommon in our application domain.

An ontology generally consists of several components that specify the domain knowledge at different levels of abstraction and from different perspectives. For the oesophagus ontology, we have distinguished between a glossary, a component that specifies the knowledge from a static perspective, i.e., a perspective in which time does not play a role, and a component taking a dynamic perspective. To ensure that no inconsistencies arise in the ontology, validity axioms have been specified for the components and their interrelationships.

The *glossary* of our ontology lists the names of the relevant domain concepts and their meanings. It serves to avoid ambiguity of terms. For example, where the domain experts use the term *metastasis* to refer to the process of conveyance of cancer cells via the blood or lymph vessels as well as to a secondary tumour resulting from this process, we have chosen an unambiguous name for each meaning and have included these, with their meanings, in the glossary.

The component that takes a *static perspective* on the knowledge addresses the hierarchical organisation of domain concepts. A concept may, for example, be a superset or a generalisation of another concept; it may also be a property or a value of another concept. This type of knowledge is represented as standard *is-a*, *part-of*, and *object-attribute-value* relations in hierarchies. Figure 1 shows a simplified fragment of the hierarchy of pathological entities (the term pathological refers to a deviation from what is normal). The component further specifies *definitional* relations that define the value of an attribute in terms of values of other attributes. For example, the attribute *T-class* of the concept *primary tumour* is defined to have the value *T3* only if the attributes *depth at site* and *depth outside site* of this concept have the values *adventitia* and *none*, respectively.

The domain of oesophageal cancer involves processes that have important effects over time. For example, the pathological process of metastasis via the blood vessels may result, in time, in a secondary tumour in the liver. We use the phrase *dynamic relation* to refer to relations between concepts that pertain to such processes. The component of our ontology that takes a dynamic perspective on the domain knowledge, specifies these relations at different description levels. At the lowest level, relations between attribute values are captured, along with their natures. This level specifies, for example, that the attribute value *presence=yes* of the process of *metastasis via blood vessels* may result in the attribute *presence* of *metastasis liver* adopting the value *yes*. At the next description level, relations between attributes are represented. They capture whether or not two attributes are related at the value level, and thus abstract from specific values. Figure 2 depicts some of the relations at this level. It shows, for example, that the pathological process of invasion may affect the depth of invasion of the primary tumour into the oesophageal wall. The tumour may further invade neighbouring organs, provided it has grown through the entire wall. We say that the depth of invasion at the primary site *enables* the invasion outside the oesophagus. Furthermore, a tumour that has invaded the oesophageal wall may *initiate* a process of metastasis via the blood vessels, which in turn may *result* in liver or lung metastases. Abstraction of the knowledge represented at the attribute level, using the *is-a* and *object-attribute* relations from the hierarchies, results in the highest description level that explicitly represents the regularities in the domain knowledge.

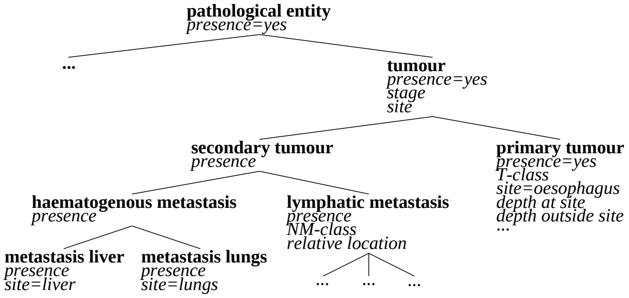


Figure 1. A fragment of the hierarchy of pathological entities

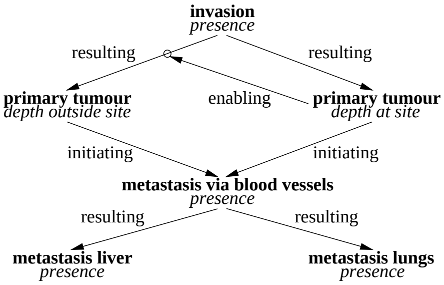


Figure 2. A fragment of the attribute-level graph of dynamic relations

5 THE DERIVATION OF THE STRUCTURE

The ontology constructed in the previous phase of our methodology is used to derive a graphical structure for a Bayesian network that captures the relevant domain knowledge. In this section, we illustrate the modelling decisions involved in deriving such a structure from the oesophagus ontology. Due to space limitations, we restrict the discussion to the knowledge that pertains to the haematogenous metastases and the depth of invasion of the primary tumour. We ignore the interrelationships with the remainder of the ontology.

We begin by selecting, from our ontology, the knowledge we would like to address. The central concepts and relations of this knowledge are then represented in several depictions. Figure 3 shows the depiction for the depth of invasion of the primary tumour. Note that the depiction describes the knowledge at the attribute level. This level often provides a convenient point of departure for deriving an initial graphical structure since attributes generally play the role of variables in a domain and can easily be associated with statistical variables. The depiction further combines relations from different components of our ontology: it includes, for example, static *definitional* relations and dynamic *resulting* relations. The *initiating* re-

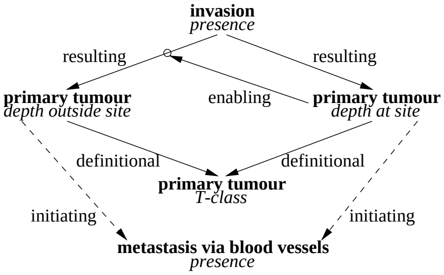
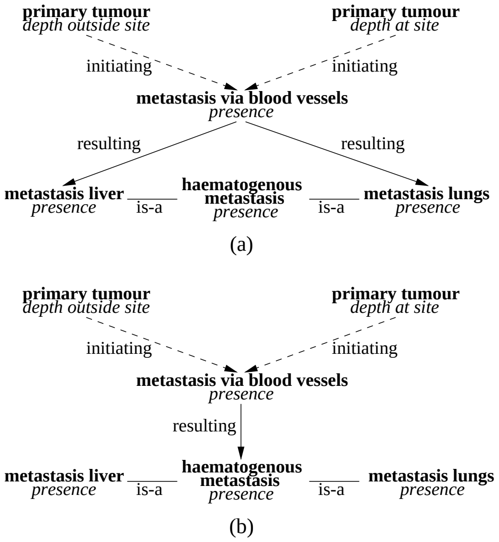


Figure 3. The depiction for the depth of invasion

lations indicated by dashed arcs connect the depictions of the two selected units of knowledge.

The knowledge that pertains to haematogenous metastases may be considered from different points of view, resulting in the two alternative depictions from Figure 4. Alternative (a) describes that the process of *metastasis via blood vessels* may result in metastases in the lungs and metastases in the liver, which are summarised as *haematogenous metastasis*. Alternative (b) captures a relation at a higher level: the process of *metastasis via blood vessels* may result in *haematogenous metastases*, which may be in the lungs or in the liver. A depiction that combines both points of view would include a redundancy and is therefore not considered. As at this stage there is no reason to prefer one alternative to the other, they will both be used to derive graphical structures. Since the *is-a* relations in the hierarchies are undirected in view of the causal interpretation that is often assigned to directed arcs, the attribute-level relations that are derived from them are left undirected in the depictions.



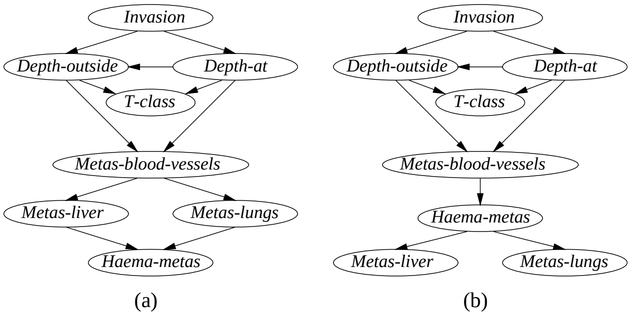
(a)

(b)

Figure 4. Alternative depictions for the haematogenous metastases

Based upon the depictions, several graphical structures are derived. For ease of presentation, we will combine the two units of knowledge into a single structure. Translating the attributes from the depictions into statistical variables is rather straightforward, because the state spaces of the attributes are exhaustive, including discrete, mutually exclusive values. Most of the arcs from the depictions are also translated directly into arcs for a graphical structure. Because the syntax of Bayesian networks only allows arcs between variables, however, translating the *enabling* relation is more involved. Since this relation in essence represents a dependence of the attribute *depth outside site* on the attribute *depth at site*, an arc is added from the variable *Depth-at* to the variable *Depth-outside*. Translating the undirected *is-a* relations further involves deciding upon a direction. The only feasible alternatives are to direct the arcs both into or both from the variable *Haema-metas*, since its relation with either location of metastases is the same. Figure 5 shows two of the resulting alternative structures; the other two alternatives are obtained by reversing the arcs between *Haema-metas*, *Metas-liver*, and *Metas-lungs*.

In the next phase, the resulting initial structures are improved and optimised. The structures are first restricted to include only variables for which probabilities can be reasonably obtained. Since the variable *Metas-blood-vessels* represents a pathological process that cannot be observed, it would be very hard for domain experts to provide probability assessments for this variable. It is therefore removed from all alternative structures, along with its incident arcs. To retain the *indirect* influences of the initiators of this process of metastasis, arcs are added from the variables *Depth-at* and *Depth-outside* to the variables *Metas-liver* and *Metas-lungs* in alternative (a), and to *Haema-metas* in alternative (b). Figure 6 shows the thus restricted graphical structures. To further optimise the structures, the variable *Invasion* and its incident arcs are removed. Because this variable will always have the value *yes* in the application under construction, it will have a *fixed* probabilistic influence on the other variables in the network. To circumvent the necessity of assessing probabilities that should not be used by the application, this fixed influence is best modelled through the assessments for the variables that are directly affected.

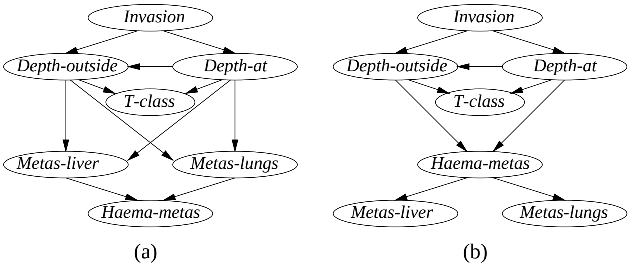


(a) (b)

Figure 5. Alternative initial structures (a) and (b)

We observe that the restricted structures specify arcs from the variable *Depth-outside* to the variables *Metas-liver*, *Metas-lungs*, *Haema-metas*, and *T-class*. These variables, however, do not depend on the specific organ being invaded, but rather on whether or not *any* organ outside the oesophagus is invaded. We note that the variable *Depth-at* has so far been defined to have the three different layers of the oesophageal wall for its values. Now, by including a fourth value that indicates that the primary tumour has grown through all the layers *and beyond*, the arcs from *Depth-outside* to the variables mentioned above are rendered superfluous. These arcs can thus be removed, thereby further reducing the complexity of the structure. Another consequence of the extension of the domain of *Depth-at* is that its relation with the variable *T-class* has become deterministic: each value of *Depth-at* is mapped onto exactly one value of *T-class*, and vice versa. We therefore replace the two variables and the arc between them by a single variable, called *Invasion-wall*. For reasons of clarity, the name of *Depth-outside* is changed to *Invasion-organs*. The results of these optimisations are shown in Figure 7.

The alternative optimised structures are now investigated as to



(a) (b)

Figure 6. Alternative restricted structures (a) and (b)

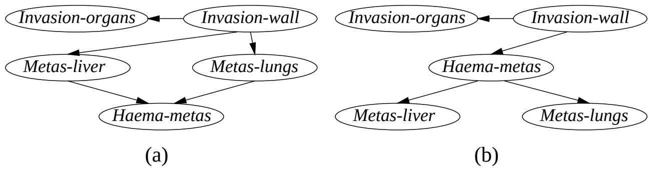


Figure 7. Alternative optimised structures (a) and (b)

whether or not they correctly capture probabilistic independence. We find that none of the alternatives faithfully represents the independences that hold in our domain. In the alternative from Figure 7(a), for example, the variables *Metas-liver* and *Metas-lungs* are shown to be independent given *Invasion-wall*. Now, if the primary tumour has just invaded the first layer of the oesophageal wall, liver metastases are not very likely. However, if it is known that there are lung metastases, then we must conclude that the process of metastasis via the blood vessels has been initiated, which renders the presence of liver metastases more likely. We therefore conclude that this independence does not hold in the domain. In alternative (b), the variables *Metas-liver* and *Metas-lungs* are shown to be independent given *Haema-metas*. This independence also does not hold in the domain: if a patient is known to have haematogenous metastases and no metastases are present in the liver, then there must be metastases in the lungs. The other independences portrayed by the two alternatives appear to hold in our domain. Investigation of the two other alternatives reveals them to be of inferior quality; they are therefore no longer considered. The representation of independence in the structures (a) and (b) can in essence be corrected by adding an arc between the variables *Metas-liver* and *Metas-lungs*. As this would increase the complexity of the structures as well as their distance from the ontology, we have chosen not to adopt this correction.

For deciding upon the final structure, we consider the nature of the inaccuracies in the representation of independence. The erroneous assumption of independence in alternative (a) pertains to the very unlikely combination of a shallow invasion of the oesophageal wall and the presence of metastases in either the lungs or the liver. Only infrequently, therefore, will the assumption be violated. The incorrect assumption of independence in alternative (b), however, is more harmful. Consider for example a patient whose test results indicate absence of lung metastases. As a consequence of the assumption of independence, the test results will erroneously be construed as being a contraindication for the presence of haematogenous metastases while the patient may in fact have metastases in the liver. This observation is confirmed by experimental results from the two alternatives, studied within the context of the entire oesophagus network using data from real patients with oesophageal cancer. Although the performance of the two networks does not differ with respect to the percentage of correctly staged patients, the probabilities with which the stages are concluded differ significantly. Especially for patients in whom the situation outlined above occurs, the network with alternative (a) concludes the correct stage with a probability that is 0.2 greater on average than the probability yielded by the network with alternative (b). Although the original network includes alternative (b), we feel that alternative (a) is the preferred alternative.

We would like to note that probabilistic independence could have been checked for the initial structures. However, checking independence is time-consuming and can be more efficiently done in a restricted structure. Moreover, a structure that correctly captures independence may become incorrect upon optimisation. For example, the initial structure from Figure 5(a) correctly represents the independence of *Metas-liver* and *Metas-lungs* given *Metas-blood-vessels*; it has become incorrect by the removal of the variable *Metas-blood-vessels*. We feel that it is advisable nonetheless to check independence as early as possible and to repeat it upon optimisation.

6 CONCLUSIONS AND FURTHER RESEARCH

Building a Bayesian network for a real-life application domain is a hard and time-consuming task that calls for the use of tailor-made knowledge-engineering methodologies. We have developed such a methodology, in which we propose to model the domain knowledge into an ontology, from which a network's graphical structure is derived in a sequence of steps. The ontology serves to document the elicited domain knowledge. The fashion in which the graphical structure is derived from the ontology provides for explicit management of modelling decisions. In the various phases, inopportune decisions are effectively forestalled. Furthermore, phase-specific guidelines support the knowledge engineer in taking the decisions required. We have illustrated the use of our methodology in the field of oesophageal cancer.

Although we have successfully applied our methodology to a real-life application domain, we are aware that it has not been fully developed and evaluated as yet. Further refinement may be needed after its use in other domains. We also expect to develop additional and more detailed guidelines to be used in deriving a graphical structure from an ontology. Future research further includes the management of uncertainties throughout a network's construction and the incorporation of procedural knowledge, to ultimately arrive at a methodology that supports not just the construction of a Bayesian network but also its embedding in a knowledge-based system.

ACKNOWLEDGEMENTS

This research has been (partly) supported by the Netherlands Organisation for Scientific Research. We thank Babs Taal and Berthe Aleman from the Netherlands Cancer Institute, who provided the domain knowledge for the construction of the oesophagus network.

REFERENCES

[1] G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. Van de Velde, B. Wielinga, *Knowledge Engineering and Management: The CommonKADS Methodology*, MIT Press, Cambridge, Massachusetts, 2000.

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

[3] M.J. Druzdzel, L.C. van der Gaag, ’Building Bayesian networks: ”Where do the numbers come from?” Guest editors’ introduction’, *IEEE Transactions on Knowledge and Data Engineering*, **12**, 481–486, (2000).

[4] K.B. Laskey, S.M. Mahoney, ’Network engineering for agile belief network models’, *IEEE Transactions on Knowledge and Data Engineering*, **12**, 487–498, (2000).

[5] M. Neil, N. Fenton, L. Nielsen, ’Building large-scale Bayesian networks’, *The Knowledge Engineering Review*, **15(3)**, 257–284, (2000).

[6] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, B.G. Taal, ’How to elicit many probabilities’, in: K.B. Laskey, H. Prade (eds), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, 647–654, (1999).

[7] G. van Heijst, A.Th. Schreiber, B.J. Wielinga, ’Using explicit ontologies in KBS development’, *International Journal of Human-Computer Studies*, **46(2/3)**, 183–292, (1997).

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

[9] M. Uschold, M. Gruninger, ’Ontologies: principles, methods and applications’, *The Knowledge Engineering Review*, **11(2)**, 93–136, (1996).