MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-259

Web-tool to Support Medical Experts in Probabilistic Modelling Using Large Bayesian Networks With an Example of Hinosinusitis

Mario A. Cypko[a], David Hirsch[a], Lucas Koch[a,b], Matthaeus Stoehr[c], Gero Strauss[d], Kerstin Denecke[a]

[a] *Innovation Center Computer Assisted Surgery, University Hospital of Leipzig, Leipzig, Germany*
[b] *Usability Department, Merseburg University of Applied Sciences, Merseburg, Germany*
[c] *ENT Department, University Hospital Leipzig, Leipzig, Germany*
[d] *International Reference and Development Center for Surgical Technology, Leipzig, Germany*

Abstract

For many complex diseases, finding the best patient-specific treatment decision is difficult for physicians due to limited mental capacity. Clinical decision support systems based on Bayesian networks (BN) can provide a probabilistic graphical model integrating all necessary aspects relevant for decision making. Such models are often manually created by clinical experts. The modeling process consists of graphical modeling conducted by collecting of information entities, and probabilistic modeling achieved through defining the relations of information entities to their direct causes. Such expert-based probabilistic modelling with BNs is very time intensive and requires knowledge about the underlying modeling method. We introduce in this paper an intuitive web-based system for helping medical experts generate decision models based on BNs. Using the tool, no special knowledge about the underlying model or BN is necessary. We tested the tool with an example of modeling treatment decisions of Rhinosinusitis and studied its usability.

Keywords:
Clinical decision support system; Bayesian network; conditional probability tables; probabilistic modelling; expert system; treatment decision; web tool; rhinosinusitis.

Introduction

For some patient-specific treatment decisions (e.g., in oncology) more patient information needs to be considered at once than a single physician is able. For this reason, time intensive meetings of experts from different medical domains (e.g., surgery, radiology and radiotherapy) are conducted to understand and discuss the entire patient situation and possible therapy options and outcomes. However, each of the experts participating in these meetings considers the patient-specific situation from an individual viewpoint or background, making a discussion and finding a unanimous decision difficult.

Intelligent clinical decision support systems (CDSSs) [1] based on probabilistic graphical models, and more specifically, a Bayesian network (BN) [2], could support physicians by modeling complex interdisciplinary treatment decisions and simulating integrated decision making. A BN's graphical model contains nodes representing information entities (IE) with a set of events that can occur (e.g. Boolean values). In the domain of medicine, IEs describe for example results regarding medical examinations, medical imaging, patient's compliance, genetic factors, or describe patient characteristics (e.g. age, gender, tobacco and alcohol consumption). Nodes are linked by directed edges. More specifically, a parental node is connected to a child node by edges representing their direct causality.

The probabilistic model is dependent on the modeled graph structure and represents the strength of the causality between an IE and its direct linked causes (parental nodes) by a conditional probability table (CPT). In a CPT, for each event of an IE probabilities are assigned based on all permutations of its parental events. Consequently, the amount of necessary probabilities for a CPT grows exponentially with the number of parental nodes and their events. A main challenge when modeling the graphical structure of BNs is to find the right balance between the granularity of IEs with their events and the complexity of the model in order to avoid large CPTs. Based on our previous work [3], CDSS for increasingly complex treatment decisions requires more detailed BN models.

The graphical model of a BN can be created by applying machine learning algorithms to a set of data collected from guidelines and other sources that define relations between IEs. The information already provide the probabilities [4-6]. Our previous study revealed a significant disadvantage of this method for complex diseases, evidences for many IEs are usually based on easier accessible and cheaper patient information (e.g. age, gender, tobacco- and alcohol consumption) [3]. Thus, proven relations as they are available in guidelines and statistics do not represent the desired natural direct dependencies in the graphical model.

In contrast to machine learning, our approach creates the model structure by considering the natural direct dependencies without limits given by statistics. The conditional probabilities are assigned in a subsequent step. In that way, the graphical model forms the basis for collecting probabilities. This kind of model-based medical evidence [7] is expensive to achieve at the moment, and manual modeling by medical experts is subjective and very time-consuming.

To overcome this limitation, we developed a web-based tool that allows experts to assign probabilities to nodes and respective events in a graphical model describing treatment decisions. In this paper, we introduce the system. It was tested on a the graphical BN model representing treatment decisions related to acute- and chronical rhinosinusits (ARS and CRS). The paper is structured as follows: In the *Methods* section, we describe the web tool to collect the CPTs for the probabilistic model. We evaluated the tool by having physicians assign probabilities to events in a treatment decision model. In the section *Evaluation and Results,* we present a comparative

assessment of the expert's probability values by using an intraclass correlation, and by analysing metadata recorded during the assignment process. Further, we present the results from a usability study. In the *Discussion* section, we describe advantages and disadvantages of the application, reasonable issues for the disparities in the expert's assessments and necessary extensions. The *Conclusion* summarizes the results and provides an outlook on future expert-based modeling in the medical domain.

Methods

Web-based Tool for Assigning Probabilities

Previous work from L.C. van der Gaag et al. [8] showed that experts can assign probabilities to IE in a given probabilistic network by formulating natural language questions and allowing the assessments to be provided on a scale. When prompted, domain experts were able to provide probabilities at a rate of over 150 probabilities per hour. Our approach is based upon that work, but goes beyond by using a computer system to digitalize the process of eliciting probabilities from domain experts through a web-based tool.

Our CPT-tool runs as a server application on the Node.js architecture and uses MongoDB as a persistence layer. The application requires a user authentication to prevent misuse of the system by unauthorized parties, and also to record user specific metadata that is used for the evaluation of the system. As a web application, the tool can be used from any browser and is thus system independent.

The CPT-tool takes as input a probabilistic network in XML format that contains all nodes of the model with its events and edges from parental nodes. We are building these networks using the open source software UnBBayes [9], but any other tool for generating BN models would be applicable. The CPT-tool automatically extracts the nodes, their events, and edges from the model to *(1)* generate an overview table and two types of questionnaires, *(2)* generate CPT specific questionnaires of permutations of parent events and *(3)* node specific selection page of probability combinations. The output of the application is again an XML file stored in the same XML format as the input, but with the probabilities set by the user.

User Interactions with the System

In the following, we describe the interactions between medical experts and the system when assigning probabilities to a graphical BN model. After logging into an account, a user can upload a probability network with events to which probabilities need to be assigned.

(1) Once the network has been analyzed by the system, the pertaining nodes are displayed in an overview table to the user, see figure 1. Each row contains the number of pages the domain expert has to complete (i.e. the number of permutations of the given node) and the number of variables each node contains. The domain expert is able to freely choose which node to start the elicitation process, by clicking the corresponding button next to the node. The required probabilities can be set at any time and from anywhere, which doesn't force the domain expert to a fixed schedule that might otherwise cause an inconvenience.

(2) The first questionnaire is similar to the elicitation sheet suggested by van der Gaag [8]: The selected IE is presented as a text fragment, but is divided into two parts. One part presents the preconditions of the current iteration, while the other part presents the events of the current node to be assessed. For example, figure 2 asks the user to assess the probability of antibiotics being used, given the state of the parent events on the left hand side. Since the events of the antibiotics node are either true or false, the user is only required to set the probability for the question: *How likely is it that antibiotics is: true.* The tool specifies the percentage bar in seven percent steps (1%, 15%, 25%, 50%, 75%, 85%, and 99%) with a text description ("(almost) impossible", "improbable", "uncertain", "fifty-fifty", "expected", "probable", and "(almost) certain") assigned to each of the percentages. However, there is still the possibility to set the exact percentage in 1% increments.

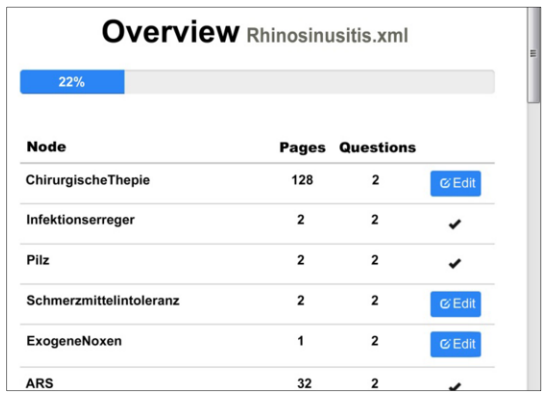

Figure 1–Overview table from CPT-tool

In case of determining the probabilities of a node with more than two events (e.g. Tumour-state), the expert is asked to set the probabilities with a total of 100%. At the end of the page, a multiple choice selection to evaluate the confidence level of the assessments is presented. The options are: very confident, confident, unconfident and very unconfident. If the assessment was unconfidently made, an additional input field is presented to describe the uncertainty of the evaluation. After submitting the probabilities, the next permutation of the node's parent events will be presented. But before, the system interpolates all probabilities to 100% in the background based on the combination of parent events, by dividing each of these probabilities by their quotient. Their quotient is the sum of the probabilities, divided by 100.

For example: The probabilities for the events $\{n_1, n_2, n_3\}$ of the node n are set with 35%, 30%, 60%. Then, the system divides each of the probability by 125%, because of (35 + 30 + 60) / 100 = 125%. 28% + 24% + 48% = 100%.

At the top of the page, the user is presented with a progress bar guiding her through the node iterations. After assessing all probabilities of a node, the user is returned to the node overview page where the recent node will be marked as completed.

(3) If a chosen node n has at least two parent nodes, before step (2), an additional questionnaire is presented to find significant combinations of parental events, see figure 3. On this page, the probability of n dependents on a non-empty and incomplete set of parent events E are to be asseessed, so that the probability of n is set for all CPT items that contain the combination of E. Significant combination of parental events are identified using the following hypothesis:

The probabilities for a node *n* are the same, if an event or a combination of events *E* of *n*'s parental nodes occurred independent on all combinations of events of all other unconsidered parental nodes of *n*. ■

The percentage for the node dependent on the selected combination is assessed by the user in the same way as in the other questionnaire (step 1). After confirming a combination with a percentage, the combination is visualized by a blue dot with a number. Same numbers indicate one combination and the number itself the order of setting the combinations.

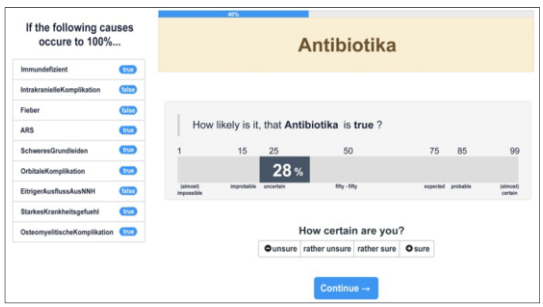

Figure 2– Example of the questionnaire from CPT-tool

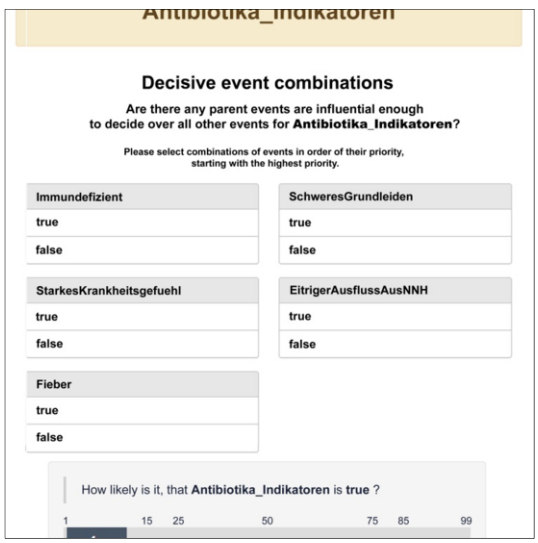

Figure 3– Questionnaire for significant combinations of parent events

Finally, after completing the assessment of all nodes, the user can export the network encompassing the elicited probabilities for each node.

Evaluation Methodology For evaluating the CPT-tool, we performed a user study with clinical experts, which also included a usibility study. Evaluation methods and results are described in the following.

Study Design: CPT-tool Used by Medical Experts

For a study of the presented CPT tool, we used a graphical model for the treatment decision of acute- and chronic rhinosinusits (ARS, CRS) comprising 75 nodes and 100 dependencies. The model was handcrafted, first by non-medical-experts based only on guidelines. For graphical representation and later inference tests, the model was implemented using UnBBayes. Three physicians were asked to correct and validate the graphical model and later, to set the CPTs by using the presented web tool. One physician was a resident physician in his second year and experienced in Bayesian theory. The other two were ENT-surgeons with 7 to 10 years of experience of ARS and CRS, with around 1000-1200 treatments a year, from this around 520 to 780 of which are surgical treatments and the remaining are conservative therapies.

The CPT-tool computed for the rhinosinusitis model contained 1526 questions, i.e., the number of probabilities that needed to be assigned. For assigning the probabilities, accounts for all three physicians were created. The experts were free to choose when and where they would use the CPT-tool to answer the questions. The only requirements were to finish a node once it was started, to keep a general overview of the node in mind, and to finish the whole assignment in two weeks. The second questionnaire for significant combinations was not part of this first study, but was assessed at a later time by resident physician.

During the assignment process, meta-data was collected to help with the evaluation of the users' decision making process. These include the previously mentioned self-evaluation of the users' choices, the amount of time it took for a user to set a probability and complete a node, and the order in which the nodes were completed. The purpose of this evaluation was to study the time required for probability assignment, to study the user satisfaction with the CPT-tool and to find out to what extent the experts agree in probability setting.

Results

Usability Study: Design and Objectives

After the first study, and based on the experiences, a usability study with 20 participants of different gender and ages between 25 and 44 years was done. Focus of the usability study was to identify potential sources of error that may lead to wrong assessments caused by the application's construction, design and usability. Therefore, a special, smaller model was created with a more intuitive example about a common topic, the accident rate of traffic, with IEs such as *road surface, the type of road, weather and season conditions*. For the tests, different usability methods were combined to get qualitative and quantitative results out of the study, such as screen capturing, voice recording, eye tracking, questionnaires and interviews. During this study, the participants were asked to think out loud and to answer questions about their actions. The participants were asked to use the CPT-tool independently by their own intuition but with the possibility to ask the usability experts if necessary.

Evaluation Results

The assignment process of the rhinosinusitis probability network took each expert an average of 7 to 12 seconds per elicitation step. At just over 1500 probabilities to assess, the experts were able to complete their task at a rate of 300 to 500 probabilities per hour, with a minimum of 2 seconds for the fastest and 26 seconds for the slowest answer. This means that a total time of 3.5 to 5 hours was necessary to complete the elicitation process for this particular model. With the aid of selecting decisive event combinations, the resident physician was able to eliminate 47.97% (n = 732) of all probabilities using less than 50 assessments.

From the metadata, we could recognize that the experts started with the nodes that had the least number of questions, before working through the remaining nodes in chronological order. Comparing the results of our experts, we found that only 148 out of the 1500 probabilities had a deviation of 50% or more. However, 501 probabilities had a deviation of over 15%. To better evaluate the assessments of the experts, we employed intraclass correlation as a form of reliability testing. The intraclass correlation coefficient kappa describes how closely the elicited values resembled each other in six categories: no agreement, slight, fair, moderate, substantial, and almost perfect agreement [10]. Figure 4 shows the kappa, on a scale of -1 to 1, for each node that has at least one parental node. Nodes without parents are ignored based on their opinionated characteristic. We did not specify in the assignment rules whether probabilities to such nodes have to be set on the basis of the specific epidemiology. The graph depicts an overall positive result, with an average kappa of 0.16, indicating "moderate agreement". Twelve of the thirty nodes depicted in the graph were categorized "substantial agreement". The few nodes found during the intraclass correlation reliability test that fall into the "fair agreement" category will be discussed and reevaluated in an upcoming post-elicitation meeting with the experts.

The two experienced experts in our study did not make use of the applications feature to report uncertainties during the elicitation process. Instead, only the resident provided feedback on several of his decisions, most frequently criticizing the epidemiology of the nodes. At this stage, it was not clear if these deviations in probabilities and the almost positive assessments of confidence originate from the differing medical expertise or is caused by the usability of the system. For this reason, we performed a usability study.

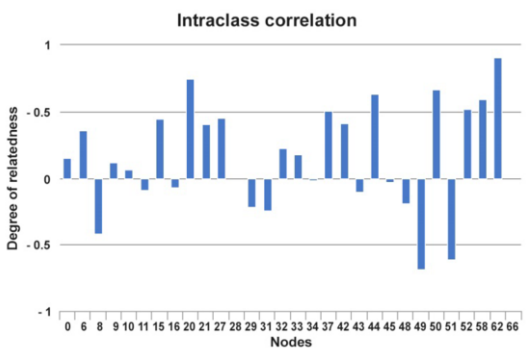

Figure 4– kappa for each node with at least one parental node from the rhinosinusitis model

In the usability study, we recognized 28 negative aspects determined by the system that we classified in three levels of severity: (1) critical – system crash or misleading to wrong probabilities (6 of 28). (2) user-unfriendly – demotivating or loss of attention (6 of 28), and (3) nice-to-have – is not affecting the rating (16 of 28). Table 1 shows the problems and the number of affected testers from the severity level 1 and 2. The additional step of presetting key combinations that was integrated only in the usability study was one of the main misleading problems. Problems with the severity level 2 lead to a longer answering time. For the most participants the study has shown first signals of demotivation after 15 to 30 minutes; after around 30 to 45 minutes participants started to loose attention. The third level is not listed as it contains subjective feelings. Only a few testers were affected by the same aspects and they did not influence the usability of the tool.

Table 1– Table of negative aspects determined by the system

Problems	# affected testers out of 20
Level of severity 1:	
Understanding of pre-setting key combinations of causes unclear	12
Group of direct causes not recognized	7
Duration of the questionnaire for one node too long	7
Only the first probability bar taken serious	5
Relation between all bars unclear	4
How to interpret the input field for uncertainty unclear	4
Some web browsers allow to stop pop-ups which prohibits functions of the tool	3
Level of severity 2:	
Comment feature hinders the assessment process	4
The node terms or subject unclear – missing instructions	3
Duration of "pages" and "questions" unclear	1
Information on lost data when closing the application	1
Program entraps to clicking through the questionnaire	1
Not enough application feedback, "unable to tell what's happening in the background"	1

Discussion

In the following , the results of both studies are discussed with examples for a better usability of the CPT-tool, and also compared with previous literature. To finish the first study with the rhinosinusits model, a follow up meeting with the experts will be needed for personal interviews and to evaluate and discuss any problems or difficulties that occurred during the elicitation process. Additionally, nodes that have led to differing assessments will be discussed to compile a convergent model for further evaluations.

The usability study with 20 participants gave highly reliable result. Early researches of Nielson and Landauer [11] showed, that five users will find about 85% of all existing usability problems and "when collecting usability metrics, testing 20 users typically offers a reasonably tight confidence interval" [12]. Based on the results from both studies, all problems with a severity level 1 should be corrected and also some of level 2 with higher numbers of affected testers. Especially, the additional step of presetting key combinations is a powerful and necessary approach for treatment decision models to minimize the large amount of probabilities. By addressing the issues revealed during the usability study, the CPT-tool can be improved to build an intuitive tool that allows domain experts to help in collecting probabilities without any need of knowledge about the underlying network or BNs. This could lead to a collection of large amount of CPTs in a short amount of time by using the crowdsourcing principle, so that later a justifiable average of probabilities can be composed.

Another significant problem is that the duration of setting the CPT for a node without interruptions is too long and leads to impatient, inattentive, fast and incorrect assessment. This problem can be solved in many ways, for example by using the average time of expert based answering, multiplied with the number of a node's questions the expert will have a better feeling for the upcoming effort. Also, the interruption during the probability setting of a node could be allowed by saving the current editing state, but with the resumption of the node its presetting history should be accessible and studied by the expert to guarantee an overview on the graph. Eliminating the problems, we expect a very intuitive tool. Another study of the upgraded application will follow.

Our CPT-tool has several advantages compared to the method from van der Gaag [8], in particular with respect to flexibility, answering rate, and precision. The separation of current and dependent events allows the user to see the preconditions all in one place, and also enables the questionnaire and underlying network to take on any form and size. With a time consumption of approximately 11 seconds per answer, the average answer rate is substantially higher. A digital solution also allows to present an exact percentage. Although it is known from van der Gaag's work that the precision of 1% increments is not necessary for the CPTs, it was one of our expert's requests to be able to distinguish between similar answers. An important and promising solution for large models is the questionnaire of significant combinations. At this stage, the usability study shows that this questionnaire is not intuitive or obvious, but it can significantly decrease the amount of questions. In a BN example with laryngeal cancer [13] some nodes in the model contain a CPT with more than ten thousands of probabilities, but can be minimized to only a small number of 20 to 30 nodes by presetting key combinations.

At the moment, the CPT-tool uses as input an XML-format given by UnBBayes's export-file that is very extensive and software specific. In the next version of the CPT-tool, this will be changed to a simpler XML-format, so that also other software export could be parsed and used in the web-tool. Additionally, an upgrade for instantiable models as multi-entity Bayesian network is planned.

Conclusion

In this paper, we introduced a web-based tool for assigning probabilities to BN models of clinical treatment decisions. The tool was evaluated with respect to usability and in a user study. The assessments we have received from the domain experts are promising even though we collected problems of three level of severity. At least level 1 problems need to be eliminated to minimize the number of wrong expert based assessments caused by the usability of the presented web-tool. The usability study not only increased the understanding of usability issues but also improved the development of an intuitive web-tool. This is a major step in encouraging medical experts to use probabilistic modeling. Based on this work, another web-tool for creating graphical models based on natural language questions is planned. These kinds of tools will allow crowdsourcing the development of CDSSs.

Acknowledgments

This research has been done within the research group "Digital Patient and Process Modeling" funded by the German Ministry of Research and Education.

We would like to thank L. Koch's usability group of the University of Merseburg for a professional and volunteered evaluation of our tool. We also like to thank Prof. Strauss's AQUA clinic team of medical experts and the International Reference and Development Center for Surgical Technology (IRDC), and also the ENT department of the University Hospital of Leipzig.

References

[1] Berner ES. Clinical Decision Support Systems: State of the Art. Agency of Healthcare Research and Quality, 2009.

[2] Pearl J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988.

[3] Stoehr M, Cypko M, Denecke K, Lemke HU, and Dietz A. A model of the decision-making process: therapy of laryngeal cancer. Int J CARS; 9, 1: 217-218, 2014.

[4] Cooper GF, and Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning; 9, 4: 309-347, 1992.

[5] Friedman N, Geiger D, and Goldszmidt M. Bayesian Network Classifiers. Machine Learning; 29, 2-3: 131–163, 1997.

[6] Getoor L, and Taskar B. Introduction to Statistical Relational Learning. MIT, 2007.

[7] Lemke HU, and Berliner L. Personalised medicine and patient-specific modelling. In Personalisierte Medizin, 14. Edition, Dresden: Health Academy; 55–164, 2010.

[8] Van der Gaag LC, Renooij S, et al. Probabilities for a probabilistic network: a case study in oesophageal cancer. Artif Intell Med 2002; 25: 123-48, 2002.

[9] Carvalho RN, Santos LL, Matsumoto S, Ladeira M, and Costa PCG. UnBBayes-MEBN: Comments on Implementing a Probabilistic Ontology Tool. In IADIS Applied Computing 2008 conference, 2008.

[10] Landis JR and Koch GG. The measurement of observer agreement for categorical data. Biometrics, 33:159174, 1977.

[11] Nielsen J, and Landauer TK. A mathematical model of the finding of usability problems, Proceedings of ACM INTERCHI'93 Conference, pp. 206-213, 1993.

[12] Nielsen J. Quantitative Studies: How Many Users to Test?. Nielsen Norman Group, 2006. Retrieved 14 December 2014 from http://www.nngroup.com/articles/quantitative-studies-how-many-users/.

[13] Cypko M, Stoehr M, Denecke K, Dietz A, and Lemke HU. User interaction with MEBNs for large patient-specific treatment decision models with an example for laryngeal cancer. Int J CARS; 9, 1, 2014.

Address for correspondence

Mario A. Cypko

Innovation Center Computer Assisted Surgery

University of Leipzig, Semmelweisstr.14, 04103 Leipzig, Germany

mario.cypko@medizin.uni-leipzig.de