ORIGINAL ARTICLE

Validation workflow for a clinical Bayesian network model in multidisciplinary decision making in head and neck oncology treatment

Mario A. Cypko[1] · Matthaeus Stoehr[2] · Marcin Kozniewski[3,4] · Marek J. Druzdzel[3,4] · Andreas Dietz[2] · Leonard Berliner[5] · Heinz U. Lemke[1]

Received: 1 September 2016 / Accepted: 25 January 2017 / Published online: 15 February 2017

Abstract
Purpose Oncological treatment is being increasingly complex, and therefore, decision making in multidisciplinary teams is becoming the key activity in the clinical pathways. The increased complexity is related to the number and variability of possible treatment decisions that may be relevant to a patient. In this paper, we describe validation of a multidisciplinary cancer treatment decision in the clinical domain of head and neck oncology.
Method Probabilistic graphical models and corresponding inference algorithms, in the form of Bayesian networks, can support complex decision-making processes by providing a mathematically reproducible and transparent advice. The quality of BN-based advice depends on the quality of the model. Therefore, it is vital to validate the model before it is applied in practice.
Results For an example BN subnetwork of laryngeal cancer with 303 variables, we evaluated 66 patient records. To validate the model on this dataset, a validation workflow was applied in combination with quantitative and qualitative analyses. In the subsequent analyses, we observed four sources of imprecise predictions: incorrect data, incomplete patient data, outvoting relevant observations, and incorrect model. Finally, the four problems were solved by modifying the data and the model.
Conclusion The presented validation effort is related to the model complexity. For simpler models, the validation workflow is the same, although it may require fewer validation methods. The validation success is related to the model's well-founded knowledge base. The remaining laryngeal cancer model may disclose additional sources of imprecise predictions.

Keywords Therapy decision support system · Bayesian network · Model validation · Laryngeal cancer · Head and neck oncology · Multidisciplinary tumor board

Introduction

A Therapy Decision Support System (TDSS) based on Bayesian networks (BN) can support multidisciplinary teams in making patient-specific therapy decisions. However, the quality of BN-based advice depends on the quality of the model. Therefore, it is vital to validate the model before it is applied in practice. In this paper, we describe a quantitative and qualitative validation workflow for a multidisciplinary cancer treatment decision in the clinical domain of head and neck oncology.

Finding the best patient-specific treatment decisions for a complex disease requires processing of large amounts of information originating from multiple sources. The ability of the human mind to handle complex and uncertain data is limited [11]. Medical experts typically deal with complexity by resorting heuristic methods focusing on a more manageable and comprehensible subset of patient information. This selection varies with experts training, specialization, background

✉ Mario A. Cypko
mario.cypko@medizin.uni-leipzig.de

[1] Innovation Center Computer Assisted Surgery, University of Leipzig, Semmelweisstr. 14, 04103 Leipzig, Germany

[2] Clinic of Otolaryngology, Head and Neck Surgery, Department of Head Medicine and Oral Health, University of Leipzig, Leipzig, Germany

[3] School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, USA

[4] Faculty of Computer Science Bialystok University of Technology, Bialystok, Poland

[5] New York Methodist Hospital, Brooklyn, NY, USA

knowledge, and experience. It may result in underestimation or disregard of certain variables and, thus, potentially lead to sub-optimal treatment decisions [11,13].

For a patient with laryngeal cancer, a treatment decision by a multidisciplinary expert team, also known as tumor board, should be standard in certified cancer centers. The goal is to adjust treatment standards to complex individual factors and to check whether it is possible to recruit the patient for clinical trials. The increasing amount of available medical knowledge and patient-specific information support this goal. They also promote a diagnostic and therapeutic variety that needs to be considered.

To support this complex decision-making process in a specific clinical setting, the Innovation Center for Computer Assisted Surgery (ICCAS) in Leipzig, Germany, is developing a clinical Therapy Decision Support System (TDSS) [4]. The TDSS is part of the Kernel for Workflow, Knowledge, and Decision Management in a Medical Information and Model Management System or MIMMS (see Fig. 1). The MIMMS allows for integration of multiple sources of data and information to facilitate Integrated (model-based) Patient Care [12].

The first clinical application to demonstrate this data integration is a treatment decision model of laryngeal cancer using a probabilistic graphical model, here specifically a BN [19]. A BN is able to combine a variety of information sources, to offer flexible and transparent decision modeling, and to provide mathematically accurate and reproducible recommendations (e.g., therapy decision, outcome, comorbidities, or quality of life) [17].

Because reasoning with BNs is based on mathematics and, hence, is theoretically correct, the quality of a model-based advice with BNs depends on the quality of the model. Therefore, validation of a BN model is the most important task after modeling and before integration of the BN into active use. Validation can be performed by a machine learning technique known as cross-validation on a data set. In case of too little data and when data are difficult to gather, expert-based validation is a necessity. An expert validation is subjective compared to data-based validation, but one could argue that it is of higher quality because the expert can study relations between variables, subnetworks, and the model behavior. However, in the BN community there are only a few reports of expert-based validation, which are conceptual or only generally described [16,18], and are all of diagnostic models.

This paper describes the construction and validation of a treatment decision model using expert knowledge supported by results from five established computer-based validation methods: accuracy, confusion matrix, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), and calibration curve.

The remainder of this paper is structured as follows. “Bayesian networks for multidisciplinary treatment decisions” section introduces Bayesian networks as well as

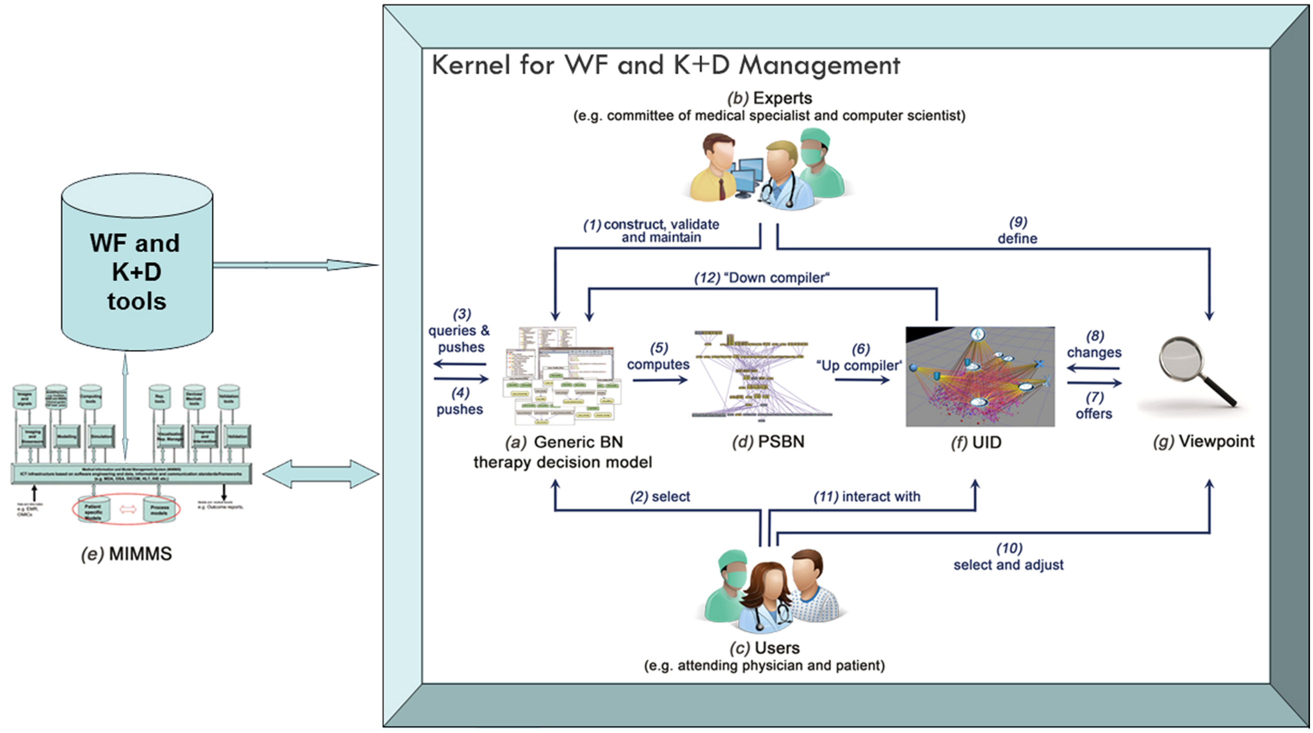


Fig. 1 Concept of a TDSS based on a generic BN [4]

our meta-structure for modeling multidisciplinary treatment decisions. "The TNM staging model of laryngeal cancer" section presents TNM staging, an appropriate subnetwork from ICCASs laryngeal cancer model for testing the validation methods. "Validation of the TNM model" section describes our validation workflow using established data- and expert-based methods. Finally, "Discussion and conclusion" section presents a general discussion and conclusion on the topic of validating complex BN models in the clinical domain of treatment decisions.

Bayesian networks for multidisciplinary treatment decisions

A BN is a probabilistic graphical model that represents a directed acyclic graph of random variables and their conditional probability distributions [17]. Specifically, a graph and the associated CPTs represent the joint probability distribution over its variables.

Each variable has a set of states, which may be Boolean valued or more detailed. Direct causal dependencies connect two directly dependent variables by a directed edge, from a parent node to its child node. The relationship between a node and its parents is quantified by a conditional probability distribution, described in a conditional probability table (CPT). In case of a variable without parents, it requires an a priori probability distribution. Once a model is created, the observations are inserted into the model, and an inference algorithm calculates the likelihood for each state of unobserved variables.

Bayesian networks have become accepted to support transparent and comprehensible clinical decision making. In the clinical context, variables may describe, e.g., diseases, symptoms, complications, and quality of life. For a variable representing the primary laryngeal cancer, states can simply be true/false, or be more specific, from T0 to T4b. In general, graphical models should reflect the causal structure of the domain, where the grade of detail should relate to the specific type of decision [6]. When multiple experts initially disagree about the model structure, they come to an agreement after some discussion [8]. CPTs should reflect knowledge from medical evidence; their parameters may be learned from data automatically, or evaluated by domain experts manually.

In case of treatment decision making, the model needs to be very comprehensive in order to reflect its complexity. We consider three types of variables to achieve a model structure that enables modeling multiple examination methods in a suitable way (see Fig. 2). The variable types are: patient situation (PN, orange nodes), examination result (EN, yellow nodes), and decision (DN, pink nodes). Directed edges represent valid causal dependencies between the three variable types. Nodes of the type examination results and decisions are observable, and the patient situation is unobservable. Examination methods have different degrees of accuracy, expressed by their CPTs.

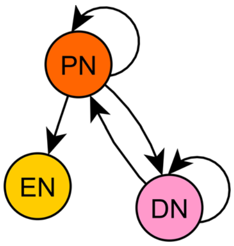


Fig. 2 Direct dependencies between three types of variables: patient situation (PN), examination results (EN), and decisions (DN)

This structure (1) allows for integrating multiple examination methods for the same type of patient information, (2) is extendable with a minimum effort in case of newly introduced medical tests or examinations, and (3) minimizes the number of direct dependencies, because examination nodes and decision nodes are only dependent indirectly.

The TNM staging model of laryngeal cancer

In case of laryngeal cancer, patient-specific treatment decisions increase both, survival rate and the quality of life. Laryngeal cancer has a worldwide annual incidence of approximately 157,000 cases with a mortality rate of around 53% [9]. However, a slight decline in mortality is observed, which results from earlier diagnoses and optimizing integrated treatment approaches [1]. In addition to the survival benefits, more attention is paid to patients' quality of life and enhancement of functional outcomes [10]. Clinical treatment guidelines in head and neck oncology, such as the National Comprehensive Cancer Network on head and neck cancers [15], provide evidence-based recommendations for the treatment of laryngeal cancer based on the TNM staging system (see "Appendix" section). The diagnostic evaluation and staging of laryngeal cancer prior to treatment typically merge into a clinical TNM staging (cTNM) that combines data from physical examination, endoscopy, and diagnostic imaging. Also, in case of a surgical treatment, a pathological TNM stage (pTNM) is defined based on (histo) pathological examinations. A pTNM is considered to be more reliable compared to a cTNM.

In collaboration between the ICCAS and the School of Information Sciences, University of Pittsburgh, for the validation analysis, we first selected a subset of the laryngeal cancer model which describes the TNM staging. This subnetwork has a sufficient complexity, is relatively well described by clinical guidelines, has an adequate evidence base, and highly impacts the patient-specific treatment decision.

The laryngeal cancer model was constructed by two ICCAS experts in a close collaboration; one was a computer scientist and the other a head and neck surgeon. Their work was supported by expert clinicians from the University Hos-

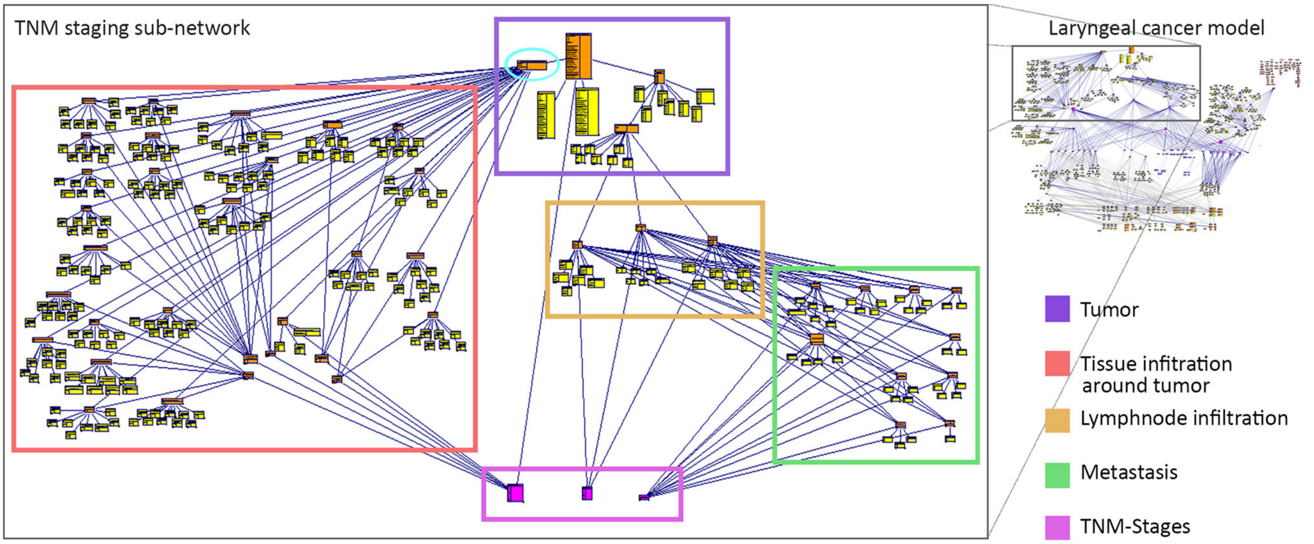


Fig. 3 The TNM staging subnetwork from the treatment decision model of laryngeal cancer

pital Leipzig (radiologists, clinical oncologists, surgeons, and radiotherapists).

The laryngeal cancer model includes variables describing the tumor physical extension (according to TNM staging), comorbidities, genetic and molecular factors, therapy options, risk factors, complications, and quality of life. The model covers the variables relevant to the tumor board, and their causal and probabilistic relationships. It encodes knowledge derived from medical guidelines and study data, as well as practical recommendations from expert clinicians in different fields. In total, the model consists of approximately 1100 variables (nodes) with more than 1300 dependencies (edges) and is described by over 1.3 million numerical parameters. Figure 3 shows the TNM staging subnetwork. In total, this subnetwork consists of 303 variables with 334 dependencies and is described by 79,815 numerical parameters. The subnetwork's parameters are assessed by a clinician using ICCAS's CPT Web tool [2], which translates mathematical equations into a natural language questionnaire. The patient dataset consists of 66 complete patient records laryngeal cancer cases. The patient records provide an average of 78 information items (ranging between 36 and 154).

Validation of the TNM model

The quality of a model can be validated by two basic approaches: quantitative and qualitative.

With the quantitative evaluation, an algorithm calculates the model accuracy automatically. For this, patient records are mandatory. Results express the model quality numerically by means of accuracy, sensitivity, specificity, etc. The number of patient records should be sufficiently large compared to the model size. Among the number of records, also the number of patient information per record has to be considered.

In practice, usually fewer cases are available, which decreases the reliability of the evaluation results. With the qualitative approach, each patient record is studied individually and interactively by a domain expert. This qualitative study is time consuming for the expert, but it provides an opportunity to check patient information in case of the model's incorrect behavior and modify the model directly [16]. In our case, the number of records was too small to obtain reliable results for the quantitative method. Please recall that we have 66 patient records and 303 model variables. However, quantitative evaluation provided misguiding predictions, which were also of interest in the qualitative studies. We applied five established quantitative methods on the TNM model given the small dataset of 66 patient records. Based on the validation results, we performed a qualitative study starting with the resulting misguiding predictions and, thereby, detected four cases of incorrect model predictions. We repeated the quantitative and qualitative validation in four major validation cycles, solving one type of issue per cycle. For validation, we used the GeNIe[1] software that supports the quantitative validation methods and tools for the qualitative validation [7]. The quantitative and qualitative validation was carried out by the clinical domain expert and the computer scientist who also built the model. The clinical domain expert evaluated the model behavior and validation results, and recommended modifications. The computer scientist operated with GeNIe, interpreted results from quantitative validation, and ensured the correct model

[1] Available free of charge for academic research and teaching use at http://bayesfusion.com/.

structure after modifications. In total, the presented validation required two intensive weeks of team work.

The following sections describe acceptable model predictions in the context of clinical treatment, present a validation workflow, introduce the five quantitative validation methods, describe the qualitative validation, and result from single modifications.

Predictions in the context of a clinical treatment decision

A prediction's acceptance depends on the decision context. A therapy decision as compared to a disease diagnosis may have a different understanding of an acceptable or unacceptable prediction.

For diagnostic models, a model's prediction may be still valuable if the correct answer (as shown in the patient record) is the second-, third-, or fourth-best prediction [16]. Furthermore, a prediction may also be valuable if the state's probability reaches a predefined probability threshold [6]. In case of cancer, for example, one might want to raise an alarm even if the probability of cancer reaches as little as 0.1.

We studied both variants of valuable predictions, separate and in combination, in the context of a clinical treatment decision. From this study, a correct answer in the second, third, and fourth probable state may be valuable in case its probability is also close to the most probable state. In contrast, this correct answer is unacceptable in case its probability is significantly lower as compared to the most probable prediction. In special cases, we call a model's node predictions misguiding, when an incorrectly predicted state has a significantly higher probability as compared to the remaining states; in clinical practice, this high confidence may exclude the consideration of other options.

We summarize the terms high confidence, uncertainty, acceptable, unacceptable, and misguiding for the following paragraphs:

(a) A model gives a *high confidence* answer about a node when the probability of the predicted state has a high probability compared to the remaining states.
(b) A model gives an *uncertain* answer about a node when the predicted state has a probability close to the probability of another state.
(c) An unpredicted correct answer is *acceptable* in case its probability is close to the probability of the predicted state.
(d) An unpredicted correct answer is *unacceptable* in case its probability is significantly lower than the probability of the predicted state.
(e) A model behavior is *misguiding* in case a wrong prediction is a certain answer.

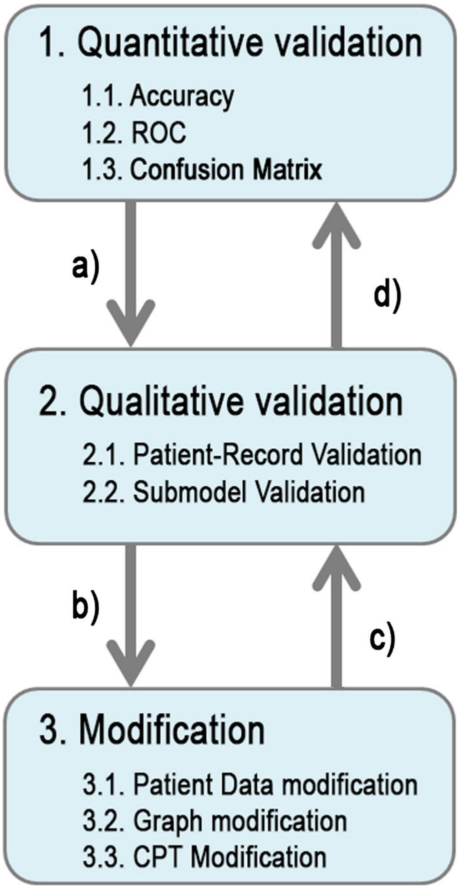


Fig. 4 Validation and modification workflow

Validation and modification workflow

The following describes the validation effort for the TNM model in more detail by a validation workflow, with an ordered list of applied methods. The process of model validation and modification consists of three steps, presented in Fig. 4; from (1) the quantitative validation through (2) the qualitative validation to (3) the modification, and back through (2) the qualitative validation to (1) the quantitative validation.

Quantitative validation is based on all available patient cases to quickly overview the model quality, and directly identify model deficiencies. One ranks identified model deficiencies from misguiding to acceptable predictions. Beginning with the misguiding predictions, from the patient cases a subset is selected that relates to exactly one of the identified model deficiencies. Based on the patient case subset, the qualitative validation is applied to study the model behavior and identify sources for the model's incorrect predictions. Sources of errors can occur in both, the model and the data. Given identified error sources, modifications are performed to solve the problem with minimum of required effort to avoid bias influences. Modifications are reviewed by repeating the qualitative validation. With a positive model behavior for the specific patient case subset, next, the quantitative validation is repeated with all patient cases to test whether the modifications develop other incorrect predictions. If also the quantitative validation is successful and our modifications are not developing new avoidable deficiencies, the validation is continued for the remaining model deficiencies discovered earlier. Otherwise, in case of wrong model behavior at the qualitative validation or identified new avoidable deficiencies at the quantitative validation, previous validation steps in order of the workflow cycle must be repeated until the problem is solved. Unavoidable deficiencies may appear after modifications, which existed before but could not be recognized given previous model deficiencies. In case of a new unavoidable deficiency, the new deficiency will be processed in a separate validation and modification cycle, and the current validation cycle is continued.

For validating and modifying a model, a selection of applied methods is necessary. We present the methods that we used in our work in Fig. 4. Quantitative validation required three of the five methods: accuracy, ROC, and the confusion matrix. Qualitative validation included testing the model behavior on patient records, followed by studying a single node or a subnetwork in more detail. Modified are first patient data, then the graph structure, and, finally, the CPT parameters. More details about specific intentions of selecting and ordering methods are described in the following sections.

Quantitative validation methods

Quantitative methods calculate a model's ability to predict values. More specific, a subset of variables called target nodes must be selected from the model. The methods exclude patient information from the test data that represent states of the target nodes, and check if the model is able to predict these values. For example, in a diagnostic setting target nodes are typically the disease nodes.

For the TNM model validation, we used five existing methods: accuracy, confusion matrix, ROC curve, AUC, and calibration curve. All methods were executed automatically using the GeNIe software. For the validation, we used the collected 66 patient records and selected as target nodes the variables that determine the TNM stages (T, N, and M state). We note that the number of collected records for a state was not balanced. For only one state per target node, we had more than 25 records; see M0, N0, and T4a. For the other states, we had on average of 7 records, for the states *TO*, *Tis*, and *N2a* we had no records at all.

Accuracy

The method *accuracy* [14] counts a model's correct predictions. Typically, a prediction is the state with the highest probability. If the prediction is equal to the patient information, it counts as correct or a hit. Accuracy is expressed by the ratio of the number of correctly predicted values to the number of all records.

The TNM model accuracy was measured in three variants, (1) considering only correct answers with highest probability, and (2) considering also correct answers with the second- and third-best predictions, and (3) accepting these second- and third-best predictions only if the distance to the most probable answer is $<32\%$. The 32% was the median of probabilities from acceptable and unacceptable predictions studying cases with correct answers in the second- and third-best prediction. Predictions were valued as acceptable and unacceptable by the domain expert.

Using the GeNIe software, we calculated the accuracy for each state of the three target nodes, T, N, and M state. From the states' accuracy, the software added up a total accuracy for each target node, and one accuracy for all target nodes. Ratios were denoted in both, a percentage and a number. However, only the most probable state predictions counted, and therefore, the second- and third-best predictions as well as the distance were calculated separately.

Confusion matrix

In case a target node has more than two states, the method accuracy summarizes the wrong predictions in one number. The *confusion matrix* [14] enables to study the details of the wrong predictions.

Confusion matrix creates for every target node a separate 2D matrix, with a node's states in rows that represent the correct answer (from patient data) against its states in columns that represent a model's predictions. The diagonal line in the matrix presents the correct predictions (*true positives*; the remaining matrix fields present incorrect predictions. The incorrect predictions in a state's row presents in numbers the model's inability to predict this state correctly (*false negatives*). The incorrect predictions in a state's column present numerically the model's inability to distinguish other states from this one (*false positives*). For this state, the remaining incorrect predictions outside its row and column are the *true positive*. From these columns, furthermore, we see the distance between wrong predictions and the expected answer.

ROC curve and AUC

With accuracy and confusion matrix, we express a model's ability to distinguish correct from incorrect predictions. *Receiver Operating Characteristic* (*ROC*) curve [20] visualizes this ability for each state separately by a curve in a 2D Cartesian coordinate system, with sensitivity against specificity. For a specific state, the sensitivity defines the *true positives* and the specificity defines *false positives*. The area under the ROC curve summarizes a ROC curve in one number running between 0.5 and 1.0, although loosing details about the curve's behavior.

Calibration curve

A calibration curve [5] expresses a model's ability to produce precise probabilities. The previous three methods (accuracy, confusion matrix, and ROC curve) enumerate the correct and incorrect predictions, but leave out the predictions' probabilities. The calibration curve is plotted in a 2D Cartesian coordinate system with probability estimates against the frequencies observed in the data. The resulting curve provides information about over- or underestimated predictions, and the distance between predicted and expected probabilities. The best case, the curve, is a straight diagonal line from (0,0) to (1,1). Below this line indicates an underestimation, and above this line an overestimation. Both underestimation and overestimation may be critical for patient outcomes; for example, an underestimation would attenuate the stage of cancer and could lead to undertreatment; an overestimation of the cancer stage may lead to an overtreatment or, in worst case, to a shift of treatment intent toward palliation, whereas a patient may survive in a curative intent. Usually, in a well-calibrated model, the higher the probabilities, the closer its calibration curve approaches the diagonal line, indicating that the model is more confident of predictions with higher probabilities. Therefore, from a curve of a state, we can discover a probability threshold of confident predictions; a predicted state with a probability above the threshold is confident to be correct.

The presented quantitative methods are all different in their informative value by means of a model's overview, details, and behavior. For the qualitative validation, we needed accuracy, then ROC curves, and the confusion matrix. AUC and calibration curve we used only to compare the initial and final model quality. The calibration curve is valuable for decision support and may have an informative value for parameter calibration. However, in this study we aimed to find and solve issues for incorrect predictions. To avoid overfitting, we did not calibrate parameters beyond this aim. The AUC results were less relevant for the qualitative validation given the ROC curves. We used accuracy for two reasons: (1) to overview the model predictions in order to select states for more detailed studies using ROC and confusion matrix and (2) to overview changes of model predictions after modifications. The ROC curve led to an understanding of the model's classification ability, unspecific in concrete predictions but quick and intuitive. From these curves, we selected states for specific details using the confusion matrix. The confusion matrix gave us the most relevant details about potential misguiding predictions.

Qualitative validation methods

To analyze the model qualitatively, experts may study both, the model behavior from patient records and direct influences through model interaction. Incorrect model predictions in the early phases of validation often indicate that the model needs some adjustment. However, a difference between model output and expert intuition is an opportunity that may lead to important insights on the part of the expert.

These studies required software with a suitable graphical user interface (GUI). We studied the model using the 66 patient records and interacted with subnetwork in case of incorrect model behavior. The results from the quantitative validation we used in order to start this qualitative validation from the misguiding predictions, because model modification also influenced the classification of the remaining TNM states.

The following sections describe a graphical user interface with functions that were important for the model interaction and validation, the patient record-based validation, the subnetwork validation, and results from model modification.

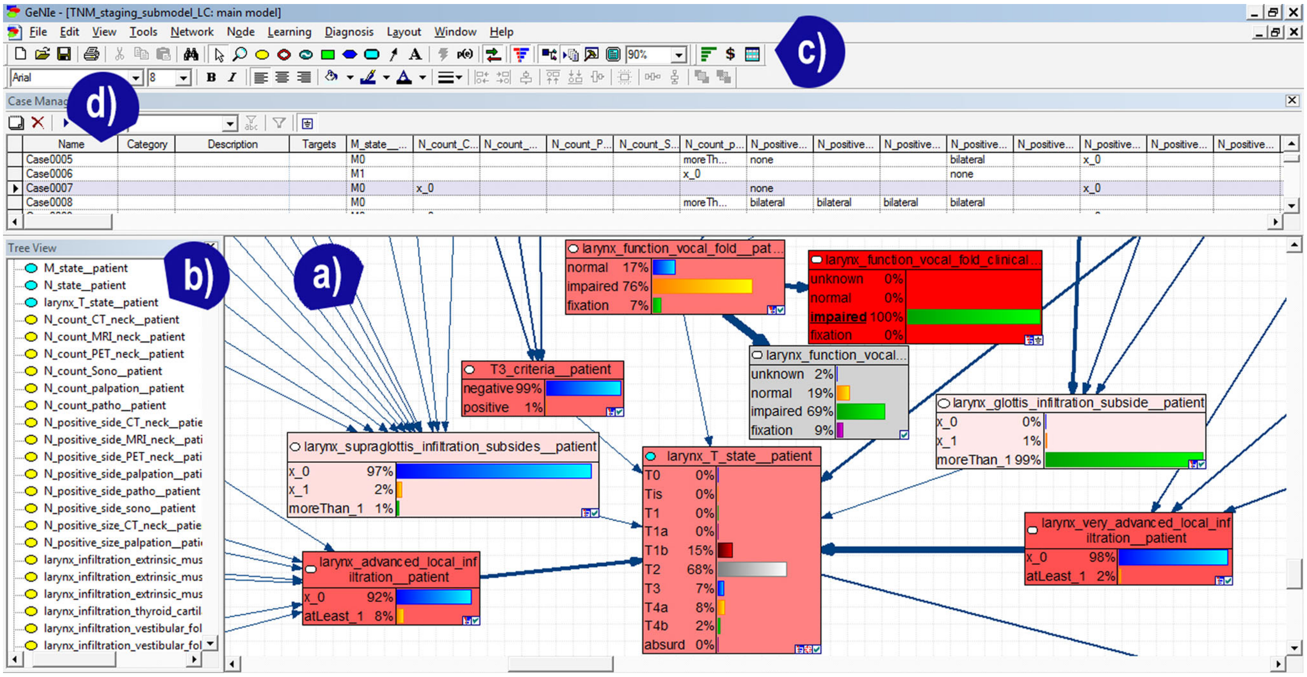
Interface and functions for expert validation

For the qualitative validation, we used GeNIe [7]. A screenshot of GeNIe's GUI is presented in Fig. 5. GeNIe supports the qualitative validation with its basic features and advanced BN functions. Basic features support interaction with both, the comprehensive model and the numerous patient records. For comprehensive models, an alphabetically ordered list of variables allows for a quick search, as well as to zoom the graph in and out, and to move across the graph structure allowing for studying subnetwork. For the patient records, a case management enables to select, modify, and save cases. Furthermore, the software updates the model inference as observations are entered or changed. Extended BN functions improve the graphical analysis. Three functions for highlighting are: strength of connections between variables based on CPTs, sensitivity analysis of target nodes based on calculated inference, and value of information based on cross-entropy which calculates the diagnostic value of observations separately for each state of target nodes.

Patient record-based model validation

An analysis using patient records allows a domain expert for studying the model behavior. From a record's number and distribution of patient information, this study gives the expert also insight into the model's reliability as a TDSS.

For the study of patient records, we processed the 66 laryngeal cancer cases one by one by loading a case into the model and analyzing the predictions and influences. In cases of incorrect model predictions, the clinician studied the graph



GeNIe - [TNM_staging_submodel_LC: main model]

Name	Category	Description	Targets	M_state_...	N_count_C...	N_count_...	N_count_P...	N_count_S...	N_count_p...	N_positive...	N_positive...	N_positive...	N_positive...	N_positive...	N_positive...	N_positive...	N_positive...
Case0005				M0					moreTh...	none			bilateral		x_0		
Case0006				M1					x_0				none				
Case0007				M0	x_0					none					x_0		
Case0008				M0					moreTh...	bilateral	bilateral	bilateral	bilateral				

File
Edit
View
Tools
Network
Node
Learning
Diagnosis
Layout
Window
Help
90%
Arial
8
Case Manag
Tree View
M_state__patient
N_state__patient
larynx_T_state__patient
N_count_CT_neck__patient
N_count_MRI_neck__patient
N_count_PET_neck__patient
N_count_Sono__patient
N_count_palpation__patient
N_count_patho__patient
N_positive_side_CT_neck__patie
N_positive_side_MRI_neck__pati
N_positive_side_PET_neck__pati
N_positive_side_palpation__pati
N_positive_side_patho__patient
N_positive_side_sono__patient
N_positive_size_CT_neck__patie
N_positive_size_palpation__pati
larynx_infiltration_extrinsic_mus
larynx_infiltration_extrinsic_mus
larynx_infiltration_extrinsic_mus
larynx_infiltration_thyroid_cartil
larynx_infiltration_vestibular_fol
larynx_infiltration_vestibular_fol
a)
b)
c)
d)
larynx_function_vocal_fold__pat...
normal 17%
impaired 76%
fixation 7%
larynx_function_vocal_fold_clinical...
unknown 0%
normal 0%
impaired 100%
fixation 0%
larynx_function_vocal...
unknown 2%
normal 19%
impaired 69%
fixation 9%
T3_criteria__patient
negative 99%
positive 1%
larynx_glottis_infiltration_subside__patient
x_0 0%
x_1 1%
moreThan_1 99%
larynx_supraglottis_infiltration_subsides__patient
x_0 97%
x_1 2%
moreThan_1 1%
larynx_T_state__patient
T0 0%
Tis 0%
T1 0%
T1a 0%
T1b 15%
T2 68%
T3 7%
T4a 8%
T4b 2%
absurd 0%
larynx_advanced_local_infiltration__patient
x_0 92%
atLeast_1 8%
larynx_very_advanced_local_infiltration__patient
x_0 98%
atLeast_1 2%

Fig. 5 Screenshot from the GeNIe software, with (*a*) the model, (*b*) a list of all variables, (*c*) functionalities to study and modify the model, and (*d*) a case manager

for finding the relevant issues. Specifically, he studied, firstly, the results of GeNIe's advanced BN features, and secondly, given these results, he went through the graph structure and studied the observations and model predictions.

Subnetwork validation

In subnetwork validation, the domain expert studies direct influences by interacting with usually small subnetworks. In detail, the expert simulates combinations of possible observations based on personal knowledge and experiences.

In this study, the clinician interacted with a node simulating observations in the node's Markov blankets. A node's Markov blanked [17] consists of node's neighboring nodes. When all nodes in a node's Markov blanked are observed, the node becomes independent of the remaining nodes in the network. Specifically, the clinicians started studying the network with the T, N, and M state nodes by simulating observations in their Markov blankets, and extended their study to Markov blankets of the neighboring nodes with incorrect influences.

Results and modifications

From the validation, we observed four problems of incorrect model predictions: (*P1*) incorrect data, (*P2*) incomplete patient data, (*P3*) outvoting relevant observations, and (*P4*) incorrect model. We decided to solve one problem at the time and test a problem's influence on the model predictions. Therefore, we solved the four problems by four modifications: (*M1*) re-staging the patients, (*M2*) including negative findings, (*M3*) adding fuzzy values, and (*M4*) modifying the model. Table 1 shows the accuracies of the T, N, and M states and the total accuracy, before and after the modifications. Finally, we present probabilities from T states predictions for analyzing the predictions' certainty (see Fig. 6).

Initially, the model accuracy based on the 66 patient cases was 76%. The patient record-based model study enabled to identify the first three problems for incorrect predictions. First, the expert (*P1*) identified 28 incorrect T, N, or M stagings in the patient data. Using retrospective data, the reasons for the mismatches were incomprehensible. The experts assumed that information may have been lost, was not updated after new examinations, or possibly of a wrong TNM staging. Finally, the clinician confirmed, after (*M1*) re-staging given the available findings, the model inferred the TNM stages in all 28 cases correctly. After correcting TNM stagings, the model accuracy increased to 89%.

Problem *P2* was caused by the fact that, in general, records consist only of positive findings and unexpected negative findings. Knowing the performed examination methods and examined body areas the clinician derived, as it is typically done in practice, the additional negative findings. Without these negative findings, the inference is performed on the CPTs, which may not fit the specific patient. We (*M2*) added the derived information to the patient records, which corrected 8% of the patient cases. However, the overall accuracy

Table 1 Accuracies of the T, N, and M states and in total, before and after the modifications of *M1* to *M4*

	Initial test	Re-staging patients (*M1*)	Including negative findings (*M1* and *M2*)	Adding fuzzy values (*M1* to *M3*)	Model modification (*M1* to *M4*)
T state	0.47	0.74	0.77	0.91	1.0
N state	0.82	0.94	0.94	1.0	1.0
M state	1.0	1.0	1.0	1.0	1.0
Total	0.76	0.89	0.9	0.97	1.0

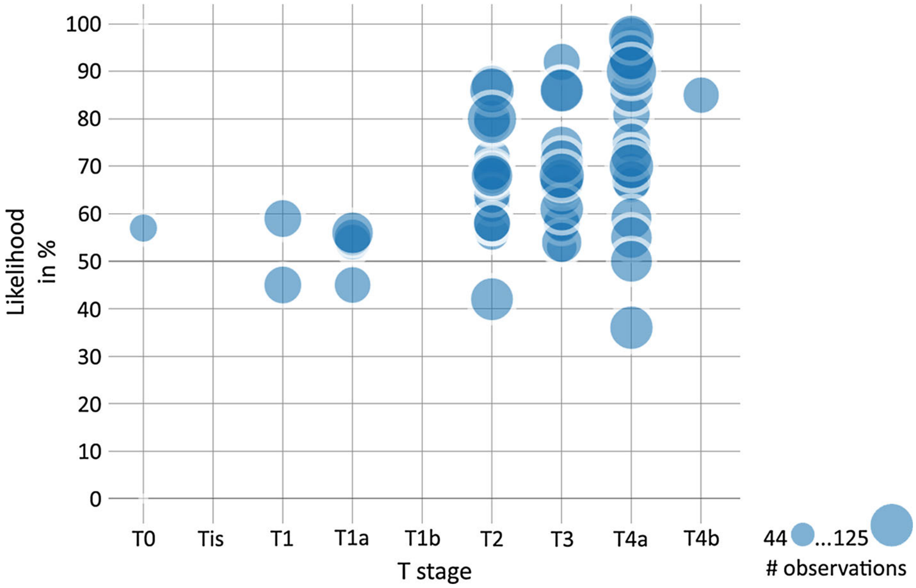


Fig. 6 A *plot* of the T states, probabilities inferred by the TNM model, and number of patient information

increased only by 1% (see Table 1), because the modification *M2* led also to problem *P3* which impaired previously correct predictions. Problem *P3* was an outvoting of results from reliable examination methods by a larger number of incorrect results from less reliable examination methods. Using GeNIe, the clinician was able to (*M3*) decrease the influence of some observations by fuzzy values, which increased the accuracy by 7% to a total of 97%.

Additionally to the patient record study, the subnetwork study discovered problem *P4*: a problem with the model detail. A variable was missing; therefore, we (*M4*) added one node and adjusted dependencies as well as edited affected CPT parameters of the new node and its children. In Fig. 3, the added node is highlighted with a circle. Finally, the total accuracy increased to 100% and the AUC of each state was at least 98%.

For a TDSS, besides the high model accuracy, also a high prediction confidence may be desired.

The T state required the most modification effort; therefore, we selected this variable to present the predictions' probability values after the last modification. Figure 6 shows a plot of T state predictions from each patient record. In this three-dimensional graph, T states are plotted against the amount of patient information and calculated probability. The probabilities reached from 36 to 97%, with an average of 71%. In general, the model predicted higher T states with more confidence. The amount of patient information was not decisive compared to the prediction confidence and T state.

Discussion and conclusion

This paper described several critically important methodologies for validating the design and structure of a TDSS based on a comprehensive BN. The utilized validation methods are well known in the machine learning community for validating models and supporting their adjustment. For both, validation and adjustment, enough test data should be made available. However, in the quantitative validation, it could be shown that even with a small set of patient records, issues can be discovered that would be faced in clinical practice, e.g., the problem with missing negative findings.

In general, for expert treatment models, the workflow described in "Validation and modification workflow" section is a valuable procedure to identify issues and find solutions. Quantitative validation provides both, an overview of model quality and details about wrongly predicted states. In turn, a qualitative validation enables to find causes of incorrect predictions as well as corresponding solutions. The presented validation effort is related to the model complexity. Multidisciplinary decision models of a similar large grade of detail will need the same validation effort independently of the represented domain. For less complex models, the validation cycle is the same, but may be simplified by using fewer methods in the qualitative validation, and correcting more issues in the same cycle. The validation workflow based on patient records and on subnetwork is always needed, but may be differently focused. Simpler models may have more patient records available and need fewer subnetwork studies.

The validation of multidisciplinary treatment models poses new challenges in data collection, modeling, and validation. These challenges should be tackled first with a clear decision as regards the role of TNM staging, which is accurately defined and well-founded. The selection of variables, setting dependencies between these variables, and also the CPT were comparatively simple, and therefore, the model was achieved with a high accuracy. This simpler subnetwork helped to identify some basic problems with modeling and validation (*M1* to *M4* in "Results and modifications" section), and should precede the validation of the more comprehensive laryngeal cancer model.

The results are promising for the clinical integration of the TNM subnetwork and for further validations of the remaining model. In case of correct model and patient data, the uncertainty in predictions is an important feedback for clinicians. The model's uncertainty may be caused by missing relevant examinations and unusual patient cases. In treatment decisions, it is common that more than one treatment option is possible. While complete certainty is often unachievable in clinical decision making, a BN provides the clinician with more certainty about the remaining uncertainty. We encourage expert validation but also point out the need for collaborative work between clinician and computer scientist to overcome the intensive validation time. The computer scientist was required for activities which, in principle, could be completely replaced by a modeling tool that is more adapted to clinicians understanding. These activities include: managing the software, interpreting the quantitative validation results, and ensuring the correct model structure with modifications.

Future developments should focus on tools to support both, BN modeling and validation [3]. A first successfully developed tool is for the assessment of conditional probabilities [2]. It is important that a validation tool follows the presented validation workflow and includes the validation methods in an abstract way that is adapted to clinicians intuitive understanding.

Acknowledgements The authors would like to thank J. Gaebel, Y. Deng, S. Oeltze-Jafra, and A. Oniśko for their valuable comments and suggestions that lead to improvements in the quality of the paper.

Funding ICCAS is funded by the German Federal Ministry of Education and Research (BMBF). The statements made herein are solely the responsibility of the authors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval For this type of study, formal consent is not required. This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent This articles does not contain patient information.

Appendix: TNM staging system for the Larynx [15]

Primary tumor (T)	
TX	Primary tumor cannot be assessed
T0	No evidence of primary tumor
Tis	Carcinoma *in situ*
T1	Tumor ≤2 cm in greatest dimension *Supraglottis:* Tumor limited to one subsite of supraglottis with normal vocal cord mobility *Glottis:* Tumor limited to the vocal cord(s) (may involve anterior or posterior commissure), with normal mobility *Subglottis:* Tumor limited to the subglottis
T1a	*Glottis:* Tumor limited to 1 vocal cord
T1b	*Glottis:* Tumor involves both vocal cords
T2	Tumor >2 cm but not more than 4 cm in greatest dimension *Supraglottis:* Tumor invades mucosa of more than one adjacent subsite of supraglottis or glottis or region outside the supraglottis, without fixation of the larynx *Glottis:* Tumor extends to the supraglottis and/or subglottis, and/or with impaired vocal cord mobility *Subglottis:* Tumor extends to vocal cord(s), with normal or impaired mobility
T3	Tumor >4 cm in greatest dimension *Supraglottis:* Tumor limited to the larynx, with vocal cord fixation, and/or invades any of the following: postcricoid area, preepiglottic space, paraglottic space, and/or inner cortex of the thyroid cartilage *Glottis:* Tumor limited to the larynx with vocal cord fixation and/or invasion of the paraglottic space and/or inner cortex of the thyroid cartilage *Subglottis:* Tumor limited to the larynx, with vocal cord fixation
T4a	Moderately advanced, local disease Lip—Tumor invades through cortical bone, inferior alveolar nerve, floor of mouth, or skin of face Oral cavity—Tumor invades adjacent structures *Supraglottis, Glottis and Subglottis:* Moderately advanced, local disease Tumor invades the outer cortex of the thyroid cartilage or through the thyroid cartilage and/or invades tissues beyond the larynx
T4b	Very advanced, local disease Tumor invades masticator space, pterygoid plates, or skull base and/or encases internal carotid artery *Supraglottis, Glottis and Subglottis:* Very advanced, local disease Tumor invades prevertebral space, encases carotid artery, or invades mediastinal structures

Regional lymph nodes (N)	
NX	Regional nodes cannot be assessed
N0	No regional lymph node metastasis
N1	Metastasis in a single ipsilateral lymph node 3 cm in greatest dimension
N2	Metastasis in a single ipsilateral lymph node >3 cm but not more than 6 cm in greatest dimension; or in multiple ipsilateral lymph nodes, none >6 cm in greatest dimension; or in bilateral or contralateral lymph nodes, none >6 cm in greatest dimension
N2a	Metastasis in a single ipsilateral lymph node >3 cm but not more than 6 cm in greatest dimension
N2b	Metastasis in multiple ipsilateral lymph nodes, none >6 cm in greatest dimension
N2c	Metastasis in bilateral or contralateral lymph nodes, none >6 cm in greatest dimension
N3	Metastasis in a lymph node >6 cm in greatest dimension

Distant metastasis (M)	
M0	No distant metastasis
M1	Distant metastasis

References

1. Chatenoud L, GaravelloW Pagan E, Bertuccio P, Gallus S, La Vecchia C, Negri E, Bosetti C (2016) Laryngeal cancermortality trends in European countries. Int J Cancer 138(4):833–842. doi:10.1002/ijc.29833
2. Cypko MA, Hirsch D, Koch L, Stoehr M, Strauss G, K D (2015) Web-tool to support medical experts in probabilistic modelling using large Bayesian networks with an example of rhinosinusitis. Stud Health Technol Inform 216:259–263
3. Cypko MA, Stoehr M, Denecke K (2015) Web-based guiding of clinical experts through the modelling of treatment decision models using MEBN with an example of laryngeal cancer. Int J CARS 10(1):163–164
4. Cypko MA, Stoehr M, Denecke K, Dietz A, Lemke HU (2014) User interaction with MEBNs for large patient-specific treatment decision models with an example for laryngeal cancer. In: Proceeding of the 28th conference for computer assisted radiology and surgery. Fukuoka, Japan
5. DeGroot MH, Fienberg SE (1983) The comparison and evaluation of forecasters. J Roy Stat Soc. Series D (The Statistician) 32(1/2):12–22
6. Diez FJ, Mira J, Iturralde E, Zubillaga S (1997) DIAVAL, a Bayesian expert system for echocardiography. Artif Intell Med 10(1):59–73
7. Druzdzel MJ (1999) GeNIe: a development environment for graphical decision-analytic models. In: Proceedings of the 1999 annual symposium of the American medical informatics association (AMIA-1999), p 1206. Washington, DC
8. Druzdzel MJ, Oniśko A, Schwartz D, Dowling JN, Wasyluk H (1999) Knowledge engineering for very large decision-analytic medical models. In: Proceedings of the 1999 annual meeting of the American medical informatics association, p 1049. Washington, DC
9. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F (2014) Globocan 2014 v1.0. Technical report on cancer incidence and mortality worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer, Lyon, France
10. Forastiere A, Weber R, Trotti A (2015) Organ preservation for advanced larynx cancer: issues and outcomes. J Clin Oncol 33(29):3262–3268
11. Kahneman D, Slovic P, Tversky A (eds) (1982) Judgment under uncertainty: Heuristics and biases. Cambridge University Press, Cambridge, England
12. Lemke HU, Golubnitschaja O (2014) Towards personal health care with model-guided medicine: long-term PPPM-related strategies and realisation opportunities within 'Horizon 2020'. EPMA J 5(1):1–9
13. Manoogian J, Benson B (2017) Cognitive bias codex. https://en.wikipedia.org/wiki/List_of_cognitive_biases
14. Moore AW, Lee MS (1994) Efficient algorithms for minimizing cross validation error. In: 11th international conference on machine learning, pp 842–846. Morgan Kaufmann, San Francisco, California
15. National Comprehensive Cancer Network (2016) Head and Neck Cancer Cancers. v1. 2016
16. Oniśko A (2003) Probabilistic causal models in medicine: application to diagnosis of liver disorders. Ph.D. thesis, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, Warsaw
17. Pearl J (1998) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, Burlington
18. Pitchforth J, Mengersen K (2013) A proposed validation framework for expert elicited Bayesian networks. Expert Syst Appl 40(1):162–167
19. Stoehr M, Cypko MA, Denecke K, Lemke HU, Dietz A (2014) A model of the decision-making process: therapy of laryngeal cancer. Int J CARS 9(Suppl 1):217–218
20. Witten IH, Eibe F (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington