Antal P., Verrelst H., Timmerman D., Moreau Y., Van Huffel S., De Moor B., Vergote I., ``Bayesian Networks in Ovarian Cancer Diagnosis : Potentials and Limitations'', in *Proc. of the 13th IEEE Symposium on Computer-Based Medical Systems (CBMS 2000)*, Houston, US, Jun. 2000, pp. 103-108., Lirias number: 178778.

Bayesian Networks in Ovarian Cancer Diagnosis: Potentials and Limitations

P. Antal *M.Sc.*[1] | H. Verrelst *M.Eng.*[1] | D. Timmerman *M.D., Ph.D.*[2] | Y. Moreau *M.Eng., Ph.D.*[1]

S. Van Huffel *M.Eng., Ph.D.*[1] | B. De Moor *M.Eng., Ph.D.*[1] | I. Vergote *M.D., Ph.D.*[2]

Abstract

The preoperative discrimination between malignant and benign masses is a crucial issue in gynecology. Next to the large amount of background knowledge there is a growing number of collected patient data that can be used in inductive techniques. These two sources of information result in two different modeling strategies. Based on the background knowledge various discrimination models are constructed by leading experts in the field, tuned and tested by observations. Based on the observations various statistical models are developed such as logistic regression models and artificial neural network models. For the efficient combination of prior background knowledge and observations, Bayesian network models were suggested. We summarize the applicability of this technique, report the performance of such models in ovarian cancer diagnosis and outline a possible hybrid usage of this technique.

1. Introduction

A reliable test for preoperative discrimination between benign and malignant ovarian tumors would be of considerable help to clinicians. It would assist them to discriminate patients for whom treatment with minimally invasive surgery or conservative management suffices versus those for whom referral to a gynecologic oncologist for more agressive treatment is needed.

One of the main goals from a medical point of view is the development of various mathematical models to predict the correct class: benign or malignant. There are two different sources of information which can be used to develop such predictive models: the biological and medical information available about the nature of the disease and the growing number of patient data. These different information sources commonly result in different types of models: medical models coming from leading experts of the field and statistical models coming from non-medical researchers. The first type of models is in general not able to exploit effectively the observations, while the second type of models is not able to exploit effectively the prior background knowledge. The Bayesian network models were suggested as a possible solution to integrate efficiently the background knowledge and observations [14]. Bayesian networks have been successfully applied in a very broad spectrum of

[1] Electrical Eng. Dept., Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94, B-3001, Leuven, Belgium. Peter.Antal@esat.kuleuven.ac.be

[2] Department of Obstetrics and Gynecology, University Hospitals Leuven. Dirk.Timmerman@uz.kuleuven.ac.be

We thank G. Fannes for his helpful comments.

applications in which the proportion of the amount of the prior background knowledge and the amount of patient data varied widely. We similarly obtained good performance comparable to the human experts and to other statistical models, however we detected certain limits of the pure application of Bayesian networks.

The paper is organized as follows. In Section 2 we give a short description of the ovarian cancer problem, the available background knowledge, the available patient data and previously suggested models. Section 3 introduces the Bayesian network models and we summarize our experiences with Bayesian network models in the ovarian cancer project. Section 4 presents the results comparing the performance of Bayesian network models to medical models based on clinical practice such as risk indices, to logistic regression models and to artificial neural network models. We conclude the paper with a summary about the advantages and limits of this technique in this project and we give a short overview of a system that tries to enhance the performance by optimally combining background knowledge and patient data.

2. The ovarian cancer problem

Ovarian malignancies represent the greatest challenge among the gynecologic cancers, and early detection is of primary importance, since currently more than two-thirds of the patients present with advanced disease.

The available abundant background knowledge is very diverse. The most common ovarian malignancies are the epithelial cancers, which arise from the cover of the ovary. Various theories hypothesize that the malignant transformation is related to the number of ovulations, to the level of gonadotropins, carcinogens and genetic deficiencies. The risk is affected by the parity (pregnancy), sterility, drug treatment for infertility, duration of lactation, oral contraceptives, foreign body (carcinogens), family history of breast and ovarian cancer, genetic deficiencies, age, age at menarche, age at menopause, hysterectomy and bilateralness. Additional measurements and observations are the following: pelvic pain, morphologic descriptors of the mass (e.g. locularility, papillation, solidness), descriptors of the vascularisation of the mass (e.g. resistance index), serum tumor markers (e.g. CA 125), echogenic descriptors of the mass, amount of ascites and the day of the cycle. While the effect of some of these variables can be reliably quantified (such as the effect of the family history and genetic deficiencies), other effects are only qualitatively known and highly subjective (such as the usage of resistance index).

In addition to the prior background information, data were collected prospectively from 300 consecutive patients who were referred to a single institution (University Hospitals Leuven, Belgium) from August 1994 till June 1997. The data collection protocol ensures that the patients have an apparent persistent extrauterine pelvic mass and excludes other causes that may have similar symptoms such as infection or pregnancy, so the primary aim is differentiation between benign and malignant masses (for a detailed description, see [16]).

Standard statistical studies indicate that a multi-modal approach, - the combination of various types of variables - is necessary for a reliable discrimination test. To assess the performance of a classifier, the Receiver Operator Characteristics (ROC) curve is used as a general measure advocated in the medical literature [17]. Previously suggested tests are based on single variables (such as CA 125, resistance index), risk indices (Lerner's scoring system, risk of malignancy index, (RMI) see e.g. [16]). Logistic regression models and artificial neural networks were similarly applied [15].

3. Bayesian network models applied to ovarian cancer diagnosis

Uncertainty is an inherent issue in nearly all medical problems. The prevailing method to manage various forms of uncertainty today is formalized within a probabilistic framework. The corresponding Bayesian statistics provides a compelling theoretical foundation that

coherent subjective beliefs of human experts should be expressible in a probabilistic framework [18]. Bayesian network models provide a practical tool to create and maintain such probabilistic knowledge bases. A Bayesian network is a knowledge model that can be used as the kernel in expert systems (for a general introduction see e.g. [12]). Furthermore the Bayesian theory describes the integration of new observations to the probabilistic model (see e.g. [13]). Consequently, the Bayesian network technique seems a natural solution to integrate prior background knowledge and data [14].

One of the main distinctive features of the discrimination task between benign and malignant masses is the centrality of the type of the mass since the data collection protocol is designed to exclude all other diseases and it ensures the presence of either a benign or a malignant mass, so every probability specification should be conditioned on the protocol. We use a single binary variable for discrimination. Taking advantage of the causality interpretation for Bayesian networks this variable can "separate" the rest of the variables into two groups: causes (such as risk factors) and effects (such as symptoms).

In the models we used the variables summarized in Section 2. The continuous and integer valued variables were discretized in accordance to the medical literature and expert knowledge. Since there are only a very restricted number of alternatives (e.g. cut-off values) we selected the prevalent discretization as shown in Table 1.

Table 1. Applied discretization schemes.

Variable	Discrete values
Age	(. ,40), [40-50), [50-60), [60-70), [70, .)
Resistance index	<0.5, 0.5<=
CA 125 serum test	<35, [35-65), 65<=
Parity	0,1,2,3,4<=

Our model building process can be separated in three different phases. In the first phase we experimented with "biological" models in which various causal models of the disease are incorporated. The specification of the structure was relatively easy, but the quantification was not possible from the literature, nor from the expert and we had a too small data set to quantify additionally introduced hidden variables. In the second phase we built "expert" models that reflect the expert's experience. The qualitative dependency-independency structure specification was again relatively easy. However the results were too biased because the medical expert participating in the project previously worked with the same collected data, so his estimates were largely based on the data set. In the third phase we built "heterogeneous" models containing biological models of the underlying mechanism quantifiable by the literature (e.g. the genetic part), parts quantified by a medical expert (e.g. age, parity distribution of the patients) and parts quantified by previously published studies (such as the effect of locularity or blood flow).

The final model, called "standard" is shown in Figure 1. For comparison we used a small and large naïve model ("small-naïve " and "large-naïve " respectively) assuming complete conditional independence between the observations conditioned on the type of the mass (i.e. two "idiot" Bayes models).

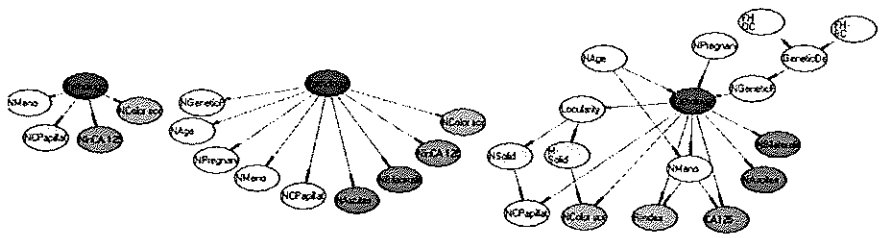

Figure 1. The "small-naïve", "large-naïve" and "standard" Bayesian networks.

The naïve models have no prior quantification (i.e. a priori specified dependency models for each variable). The sources of quantification of dependency models in the standard model are described by Table 2.

Table 2. Sources of quantification for the "standard" model.

Variable	Source of quantification
Family members with ovarian cancer	Expert
Family members with breast cancer	Expert
Genetic risk	Expert
Genetic deficiency	Literature [1,2,3,4,5], expert
Pregnancy	Expert
Age	Expert
Pathology	Literature [1,2,3,4,5,6,7], expert
Menopausal status	Expert,
Locularility	Literature [8]
Color score	Literature [9]
Resistance index	Literature [10]
Bilaterality	Literature [6]
Ascites	Expert
Papillation	Literature [8]
CA 125	Literature [11]

Because of the extensive and complex usage of the prior knowledge we used a strict documentation method to track the route of the prior information from studies into the model. Conversion formulas were constructed to compile the raw prior knowledge to be compatible with the conditions of the task and the format of the Bayesian network. The following list contains the high-level steps of this process:

1. Make a list of all prior knowledge about variables, discretizations, existing dependency models, etc.
2. Classify different types of priors that exist (from exactly specified prior sub-models to high level guesses about qualitative dependencies).
3. Select a "coverable variable set" what seems to be quantifiable from the prior background knowledge and the available data.
4. Specify a complete domain model by following the standard construction mechanism for Bayesian networks and considering the existing prior sub-models.
5. Construct secondary conversion models and formulas to quantify the final model (incorporating hyperparameters about confidence, conditioning on the conditions of the discrimination task, etc.).
6. Quantify, documenting the sources of the information for interpretation, modification and maintenance.

4. Results

Five Bayesian network models are investigated: the two naïve models and the standard model in three contexts: prior quantification without hyperparameter update (i.e. learning), update from the data without prior quantification and prior quantification with hyperparameter update. The performance is assessed w.r.t. the area under the ROC curve computed by exact trapezoidal integration (the standard error (SE) was computed following [17]). Additionally, Table 3 contains the sensitivity, specificity, positive predictive and negative predictive values (in percentage).

The hyperparameters in the Bayesian networks are updated using 75% of the data set,

the rest is used as a test set for estimating the area under the ROC curve (averaged over 1000 cross-validation sessions).

The risk of malignancy model is based on the menopausal status, CA 125 serum test and on a morphologic score. The logistic regression model, the artificial neural network model [16] and the "small-naïve" Bayesian network model have the same four inputs: menopausal status, CA 125 serum test, color score and papillarities.

Table 3. Performance of models

Model	ROC (%)	SE(%)	Sens.	Spec.	PPV	NPV
Serum CA 125 (U/ml) [16]	87.4	3.4	79.6	81.5	62.9	91.0
Risk of malignancy index [16]	89.1	3.2	87.8	74.2	57.3	93.9
Logistic regression [16]	90.4	6.0	85.7	81.1	63.2	93.8
Artificial neural network [16]	95.1	3.9	87.5	92.7	82.4	95.0
Small-Naïve (BN)	93.1	3.9	94.7	74.1	84.2	90.0
Large-Naïve (BN)	93.8	3.7	96.6	79.9	90.1	92.6
Standard-prior-no-update (BN)	90.4	2.3	93.6	72.3	81.1	89.8
Standard-no-prior-update (BN)	95.0	3.4	94.2	83.5	84.8	93.8
Standard –updated-prior (BN)	95.2	3.4	94.7	83.4	86.1	93.7

The previously reported results were achieved on a smaller data set under different testing conditions [16]. Comparison of the results in Table 3 considering this difference shows that the Bayesian network models have a similar performance as the best performing artificial neural network model. The performance of the Bayesian network models are significantly better than the RMI and CA 125 (for significance testing [17] was used). Although the "standard" prior Bayesian network without update has a slightly better performance than the RMI, the difference is not significant.

5. Conclusion and future work

Our experience confirms that Bayesian networks provide a practical solution for representing medical knowledge, performing inferences, and learning. They are particularly effective in integrating various prior sub-models together. For example, the "standard" Bayesian network model (quantified from previous studies about various sub-parts of this model) has the same performance as the RMI score, which is the accepted reference method. The negative effect of the crude discretization scheme is compensated by the multi-modal approach in an appropriate dependency structure. Additionally, the prior quantification is dominated by a sample containing 50 to 100 random patient cases, which means that in our case (300 patients) the prior has no effect on the final performance. An important advantage of such "white-box" models is that they can be used for explanation or a semantic sensitivity analysis. Furthermore, because of their decomposed nature, they can be extended to perform finer sub-classification, which is the next phase of our project.

However for an efficient integration, the management of heterogeneous types of prior information needs better support (such as the enumeration, conversion and documentation of various types of not-formalized prior information). Another bottleneck is the discretization: Indeed, it is usually fixed by expert knowledge (e.g., choice of various cut-off values) and the conversion of these frequently incompatible schemes to a better discretization scheme or to a continuous scale often wastes a lot of prior knowledge. The same problem arises with respect to the structure, since the prior knowledge frequently consists of multiple structures.

These and other theoretical constraints force us to combine the advantages of the Bayesian network models (understandable knowledge representation) and black-box models (efficiently learnable representation) instead of pitting them against each other. We are currently testing such a hybrid methodology and will report about it in a later publication.

Acknowledgements

I. Vergote is full professor, D. Timmerman is associate professor at the University Hospitals Leuven. B. De Moor is a Research Associate at the F.W.O. (Fund for Scientific Research - Flanders) and professor extra-ordinary at the K.U.Leuven. S. Van Huffel is a Senior Research Associate with the F.W.O. P. Antal is a Research Assistant with the K.U.Leuven. H. Verrelst is a Research Assistant with the I.W.T. (Flemish Institute for Scientific and Technological Research in Industry). This work was carried out in collaboration with Data4s Future Technologies NV at the ESAT laboratory and supported by grants and projects from the K.U.Leuven Interdisciplinary Research Program (IDO) on ovarian tumor classification; the Flemish Government: Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) and F.W.O. project G.0262.97: Learning and Optimization: an Interdisciplinary Approach and the F.W.O. Research Communities: ICCoS (Identification and Control of Complex Systems) and Advanced Numerical Methods for Mathematical Modelling and Bilaterale Wetenschappelijke en Technologische Samenwerking Flanders - Hungary, Project BIL96/19; from the Belgian State, Prime Minister's Office - Federal Office for Scientific, Technical and Cultural Affairs - Interuniversity Poles of Attraction Programme (IUAP P4-02 (1997-2001): Modeling, Identification, Simulation and Control of Complex Systems. The scientific responsibility is assumed by its authors.

References

[1] J. S. Berk: Practical Gynecologic Oncology 2nd Ed., Williams and Wilkins
[2] D. F. Easton et al., Breast and Ovarian Cancer Incidence in BRCA1-Mutation, Am. J. Hum. Genet, 56, 1995, pp. 265-271
[3] A. A. Langston et al., Hereditary ovarian cancer, Gyn. Onc. And Path., Rapid Science Publishers, 9, 1997,pp. 3-7
[4] A. S. Whittemore et al., Prevalence and Contribution of BRCA1 Mutations in Breast Cancer and Ovarian cancer, Am. J. Hum. Genet, 60, 1997, pp. 496-504
[5] E. B. Claus et al., Autosomal Dominant Inheritance of Early-Onset Breast Cancer, Cancer, vol. 73, no. 3, 1994, pp. 643-650
[6] SEER Cancer Data, National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) program (US), 1998
[7] A. S. Whittemore et al., Characteristics Relating to Ovarian Cancer Risk (I, II, III, IV), Am. J. Epidemiol. 136, 1992, pp. 1175-1220
[8] S. Granberg et al., Macroscopic characterization of ovarian tumors and the relation to the hitological diagnosis, Gynecol Oncol, 35, 1989, pp. 139-144
[9] L Valentin, Gray scale sonography, subjective evaluation of the color Doppler image and measurement of blood flow velocity for distinguishing benign and malignant tumors of suspected adnexal origin, Eu. J. Obst.&Gyn. and Repr. Bio.72, 1997, pp. 63-72
[10] A. Tekay et al., Validity of pulsatility and resistance indices in classification of adnexal tumors with transvaginal color Doppler ultrasound, Ultrasound Obstet. Gynecol, 2, 1992, pp. 338-344
[11] N.J. Finkler et al., Compraison of Serum CA 125, Clinical Impression, and Ultrasound in the Preoperative Evaluation of Ovarian Masses, Obstretics&Gynecology, vol. 72, no. 4. Oct. 1988, pp. 659-663
[12] E. Castillo et al., Expert systems and probabilistic network models, Springer 1997
[13] M.J.Jordan(ed), Learning in graphical models, Kluwer, 1999
[14] D. Heckerman. et al.: Learning Bayesian networks: The Combination of Knowledge and Statistical Data, Machine Learning, 20, 1995, pp. 197-243
[15] D. Timmerman et al., Artificial neural network models for the pre-operative discrimination between malignant and benign adnexal masses. Ultrasound Obstet Gynecol 1999; 13: 17-25.
[16] D.Timmerman, Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology, Ph.D. dissertation, Leuven University Press, 1997
[17] J. A. Hanley et al., The Meaning and Use of the Area under Receiver Operating Characteristic (ROC) curve, Radiology, 143, April 1982, pp. 29-36
[18] J. M. Bernardo et al., Bayesian theory, John Wiley&Sons, 1994