Human Reproduction Vol.21, No.7 pp. 1824–1831, 2006 doi:10.1093/humrep/del083
Advance Access publication April 6, 2006.

Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression

O.Gevaert[1,6], F.De Smet[1,2], E.Kirk[3], B.Van Calster[1], T.Bourne[3], S.Van Huffel[1,4], Y.Moreau[1], D.Timmerman[5], B.De Moor[1] and G.Condous[3]

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg, Leuven, [2]Medical Direction, National Alliance of Christian Mutualities, Haachtsesteenweg, Brussels, Belgium, [3]Early Pregnancy, Gynecology Ultrasound and MAS Unit, St George's Hospital Medical School, London, UK, [4]University Center for Statistics (UCS), Katholieke Universiteit Leuven and [5]Department of Obstetrics and Gynecology, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, Belgium

[6]To whom correspondence should be addressed at: Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. E-mail: olivier.gevaert@esat.kuleuven.be

Availability of models: http://homes.esat.kuleuven.be/~bioiuser/PUL

BACKGROUND: As women present at earlier gestations to early pregnancy units (EPUs), the number of women diagnosed with a pregnancy of unknown location (PUL) increases. Some of these women will have an ectopic pregnancy (EP), and it is this group in the PUL population that poses the greatest concern. The aim of this study was to develop Bayesian networks to predict EPs in the PUL population. METHODS: Data were gathered in a single EPU from all women with a PUL. This data set was divided into a model-building (599 women with 44 EPs) and a validation (257 women with 22 EPs) data set and consisted of the following variables: vaginal bleeding, fluid in the pouch of Douglas, midline echo, lower abdominal pain, age, endometrial thickness, gestation days, the ratio of HCG at 48 and 0 h, progesterone levels (0 and 48 h) and the clinical outcome of the PUL. We developed Bayesian networks with expert information using this data set to predict EPs. RESULTS: The best Bayesian network used the gestational age, HCG ratio and the progesterone level at 48 h and had an area under the receiver operator characteristic curve (AUC) of 0.88 for predicting EPs when tested prospectively. CONCLUSIONS: Discrete-valued Bayesian networks are more complex to build than, for example, logistic regression. Nevertheless, we have demonstrated that such models can be used to predict EPs in a PUL population. Prospective interventional multicentre studies are needed to validate the use of such models in clinical practice.

Key words: Bayesian networks/classification/ectopic pregnancy/pregnancy of unknown location/prior information

Introduction

The more liberal use of home pregnancy tests and access to early pregnancy units (EPUs) have led to an increase in the number of women undergoing early transvaginal ultrasound scans (TVS) to locate, date and confirm viability of their pregnancy. This approach has resulted in more women being diagnosed with a pregnancy of unknown location (PUL) or inconclusive scan (an empty uterus and no adnexal mass on TVS). This group of women is defined as having a positive pregnancy test and no sign of a pregnancy on TVS. Within the PUL group, there are four clinical outcomes: a failing PUL, an intrauterine pregnancy (IUP), an ectopic pregnancy (EP) or a persisting PUL. A persisting PUL is defined as a case where there is a plateauing in serum HCG levels without visualization of an IUP or EP. Persisting PULs are rare and have therefore been excluded from this analysis. The location of the failing PULs remains unknown and consists of failing IUP and failing EPs. A major challenge is the detection of EPs within this PUL group. Most PULs are non-EPs (Condous *et al*., 2004a), and therefore the most important diagnostic problem is the correct classification of the EPs.

The EPs within this PUL population pose the greatest risk—they are the most important cause of maternal death in early pregnancies (Lewis and Drive, 2004). Previously, the diagnosis of EPs was based on the classical triad of amenorrhoea, lower abdominal pain and vaginal bleeding. This is in contrast to the current situation where women present earlier with less specific symptoms. Various thresholds based on serum HCG and progesterone levels have been proposed resulting in cut-off models (Kadar *et al*., 1981; Condous *et al*., 2002). Other methods have included a single-visit strategy and also the subjective interpretation of biochemical data by an expert (Condous *et al*.,

2004b, 2005a). These methods were capable of detecting failing PULs and IUPs, but they produced too many false-positives. This justified the use of mathematical modelling techniques to predict EPs.

Recently, logistic regression has been applied to predict the outcome of PUL (Condous *et al*., 2004a) with satisfactory results for predicting EPs. Bayesian networks provide a more complex framework than logistic regression and allow arbitrary relations between all the variables (Pearl, 1988; Neapolitan, 2005). Moreover, a Bayesian network is a white-box model and allows the incorporation of expert knowledge into the model at various levels. Using prior information from experts, we can direct the model-building process and combine this information with the data.

The aim of this study was to evaluate the use of discrete-valued Bayesian networks in combination with different forms of prior information when predicting the outcome of PULs. Previous work (Condous *et al*., 2004a; Van Calster *et al*., 2005) demonstrated that predicting failing PULs and IUPs is very accurate, whereas predicting EPs in a PUL population is challenging. Therefore, we concentrated on the use of discrete-valued Bayesian networks to discriminate EPs from non-EPs. The results are compared with logistic regression (Condous *et al*., 2004a) for the prediction of EPs.

Materials and methods

Data collection

Data were collected prospectively from consecutive women who presented with a PUL, at St George's Hospital, London between June 2001 and October 2004. All women underwent TVS using a 5-MHz probe (Aloka SSD 900, 2000 or 4000, Keymed, Southend, UK and Aloka, Tokyo, Japan). Blood was taken to measure the levels of serum HCG (World Health Organization, Third International Reference 75/537) and progesterone (Roche Elecsys 2010 Progesterone II test) using automated electrochemiluminescence immunoassays (ECLIAs). These levels were measured at 0 and 48 h. The data set contained 10 variables (Table Ia and b) among which were both discrete and continuous variables. The discrete variables were (i) the presence of vaginal bleeding, (ii) the presence of free fluid in the pouch of Douglas, (iii) the character of the midline echo (intact or disrupted) and (iv) the presence of lower abdominal pain. The continuous variables were (i) age, (ii) the thickness of the endometrium (in mm), (iii) the number of gestation days, (iv) the HCG ratio defined as the HCG at 48 h (in U/l) divided by the HCG at 0 h and (v) the progesterone level at 0 and 48 h (in nmol/l). The continuous variables were discretized according to discretization intervals specified by an expert in early pregnancy. This was possible because this expert based the intervals on past experience and chose thresholds that were empirically known to reflect clinical states. Moreover, these discretization intervals were specified by balancing two issues, keeping as much information as possible while limiting the number of intervals to reduce the number of parameters.

The women were followed-up until a final diagnosis could be established: a failing PUL, an IUP or an EP. A failing PUL was confirmed when there were persistent negative sonographic findings in the presence of falling serum HCG levels, ultimately reaching the detection level (i.e. lower than 5 U/l). An IUP was confirmed sonographically during follow-up with the presence of a gestational sac eccentrically placed within the endometrial cavity. The diagnosis of an EP was based on the positive visualization of an adnexal mass (Condous, 2005b).

Ultrasonographic diagnosis of an EP was based on the following grey-scale appearances: (i) an inhomogeneous mass adjacent to the ovary and moving separate to this—we have called this the blob sign (Condous *et al*., 2005b); or (ii) a mass with a hyper-echoic ring around the gestational sac referred to as the bagel sign (Condous, 2005b) or (iii) a gestational sac with a fetal pole with or without cardiac activity

Table I. Descriptive statistics of the model-building and prospective data set for the continuous variables (a) and discrete variables (b) separately for ectopic and non-ectopic pregnancies

Variable name	Model-building data set		Validation data set	
	Non-ectopic	Ectopic	Non-ectopic	Ectopic
(a) Continuous variables: mean (minimum–maximum)				
Age	30 (15–48)	30 (19–38)	30 (15–49)	30 (22–39)
Endometrial thickness (mm)	11 (1.5–35)	11 (2.5–26)	11 (2–31)	11 (3.8–22)
Gestation days (days)	42 (13–100)	44 (23–85)	43 (10–93)	42 (19–93)
HCG ratio [HCG 48 h (U/l)/HCG 0 h (U/l)]	1.3 (0.11–4.8)	1.3 (0.33–3.1)	1.2 (0.08–4.2)	1.3 (0.34–2.4)
Progesterone 0 h (nmol/l)[a]	20 (1–190)	26 (3–191)	10 (1–191)	3 (4–89)
Progesterone 48 h (nmol/l)[a]	10 (1–190)	24 (2–178)	6 (1–250)	22 (5–84)
(b) Discrete variables: *n* (%)				
Bleeding				
No	260 (47)	14 (32)	100 (43)	9 (41)
Yes without clots	224 (40)	25 (57)	101 (43)	13 (60)
Yes with clots	71 (13)	5 (11)	34 (14)	0 (0)
Free fluid				
No	450 (81)	36 (82)	198 (84)	20 (91)
Yes	105 (19)	8 (18)	37 (16)	2 (9)
Midline echo				
Intact	475 (86)	37 (84)	205 (87)	18 (82)
Disrupted	80 (14)	7 (16)	30 (13)	4 (18)
Pain				
No	294 (53)	26 (59)	119 (51)	15 (68)
Yes	261 (47)	18 (41)	116 (49)	7 (32)

[a]In this case, we used medians, because the standard deviation was in some cases larger than the mean.

(Condous, 2005b). The diagnosis was subsequently confirmed at laparoscopy with histological confirmation of chorionic villi in the Fallopian tube in those women who underwent surgery. If an EP was not visualized, but there was a high index of suspicion based on symptomatology, clinical findings and suboptimal rises of serial serum HCG levels, a laparoscopy was performed with or without an evacuation of the uterus.

Before starting the model building, the data set was randomly split into a set that we called the 'model-building data set' and a set that we called the 'validation data set'. The splitting was done in a stratified manner to ensure that the proportion of EPs in both sets was essentially the same.

All hypothesis tests were two-sided, and the level of significance was 0.05.

Bayesian networks

Bayesian networks were applied to detect the EPs in the PUL population. A Bayesian network is a probabilistic model (Pearl, 1988) that consists of two parts: a dependency structure and local probability models. The dependency structure specifies how the variables are related to each other by drawing directed edges between the variables without creating any directed cycles. Usually, a variable only depends on a few other variables, called the parents. The second part of this model, the local probability models, specifies how the variables depend on their parents. We used discrete-valued Bayesian networks, which means that these local probability models can be represented with conditional probability tables (CPTs). Such a table specifies the probability that a variable takes a certain value, given the value of its parents. Figure 1 shows an example of a small Bayesian network with three variables and the corresponding CPTs.

Learning Bayesian networks

Learning discrete-valued Bayesian networks from data proceeds in two steps. First, the dependency structure that best explains the data is constructed. This is done using a heuristic search strategy combined with a scoring metric (Cooper and Herskovits, 1992; Heckerman *et al*., 1994). The scoring metric reflects the fit of the structure to the data, possibly combined with prior knowledge. The best dependency structure is found by iteratively making all small changes to the current dependency structure and adopting the change with the highest score. Because of the greedy nature of this search strategy, it is likely to come up with a suboptimal model, and therefore the model-building process is repeated several times with different initialization parameters. Then, the model with the highest score is selected. The second step consists of estimating the parameters of the local probability models for the selected model. This amounts to filling in a CPT for every variable and every possible value of its parents using the data.

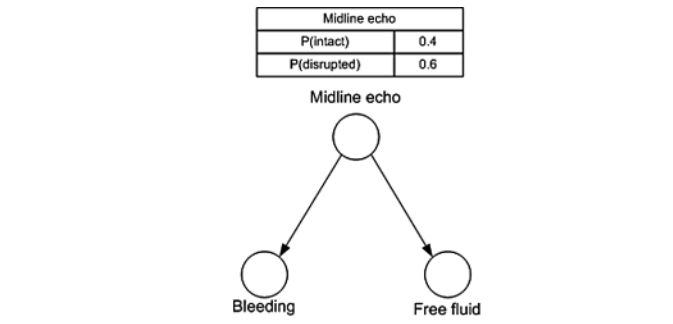

Figure 1. A small Bayesian network with three variables and the corresponding conditional probability tables (CPTs). The midline echo node has no parents; therefore, the CPT is a special case and contains the prior probability for the different values of this variable. The other variables, bleeding and free fluid, have midline echo as their parent and have a separate probability distribution for each value of their parent. This corresponds to the columns in the CPTs of bleeding and free fluid.

Expert prior information

The building of Bayesian networks allows the incorporation of prior knowledge in combination with the data. In this case, the prior knowledge was gathered from an expert in the field of early pregnancy. It was included in the model building during structure learning in two ways: at the structure level and at the parameter level. This was done by specifying both a prior distribution for the structure (the structure prior) and a prior distribution for the parameters (the parameter prior). Our expert first specified the structure prior, which yields the probability that an edge will occur in the dependency structure. In Figure 1, for example, this means that we have to specify the probability for each directed edge between all combinations of two variables. Because there are three variables, there are six possible edges, and therefore six probabilities have to be specified (between two variables there are two possible edges, one in each direction). After assessing these prior probabilities, this results in a table that specifies the probability that a directed edge occurs from one variable to another. Our expert assessed these prior probabilities between all the variables in our data set. Figure 2 shows the result of this process graphically by increasing the thickness of the edges when the prior probability increases. Using this information, the structure prior can guide the structure learning by estimating the prior probability of a structure based on the knowledge of an expert.

Next, our expert specified a parameter prior. The parameter prior consists of a complete model (the dependency structure and the corresponding local probability models) based on past experience. This amounts to first specifying the structure of a Bayesian network; usually this is based on assessing causal relations between the variables. Figure 3 shows the structure that was specified by our expert. The second step consists of specifying a probability table for each variable in Figure 3 similar to the probability tables that are shown for the Bayesian network in Figure 1. This is only feasible when the number of variables is low and when the relationships between the variables are sparse. In this case, the number of probability tables and their size are small. The parameter prior is mathematically equivalent to introducing extra cases according to the subjective knowledge of an expert. Both priors influence structure learning via the scoring metric (Heckerman *et al*., 1994).

Model-building methodology

Selection of the optimal use of prior information

We evaluated the performance of the different combinations of prior information (no priors, structure prior, parameter prior and both priors) on the model-building data set. This was done by randomizing the model-building data set 100 times for each combination of prior information, in a stratified way, into a set of 70% of the patients used to build the model and a set of 30% to estimate the area under the

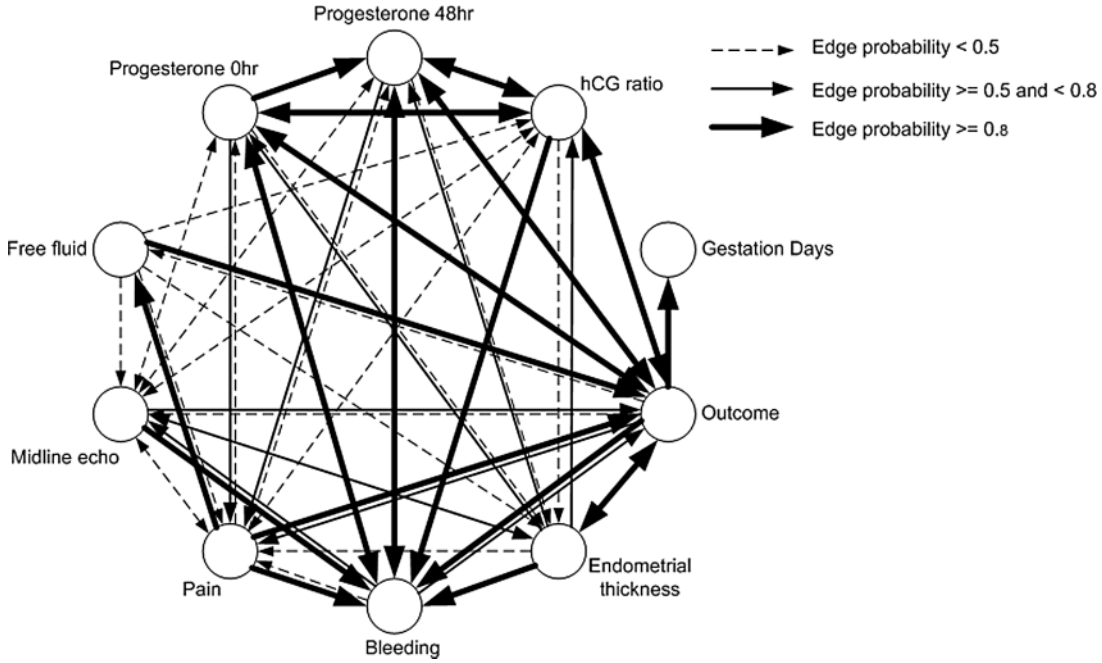

Figure 2. Visualization of the structure prior. The lines represent the structure prior where the probability was larger than zero as specified by an expert. The thickness of the edges is proportional to the probability that the edge will occur according to the subjective knowledge of the expert.

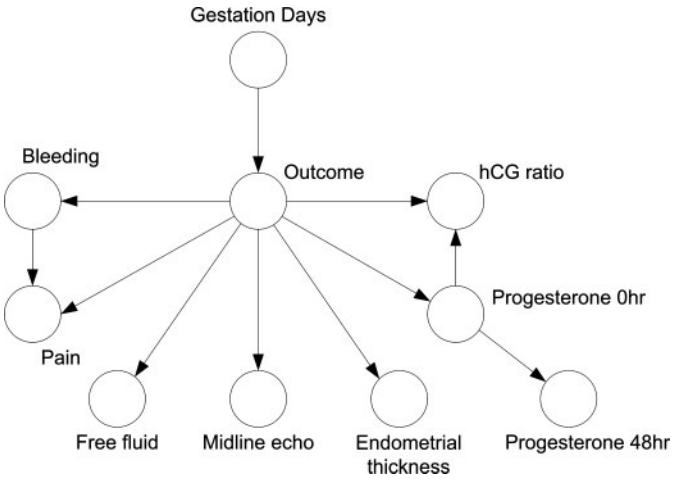

Figure 3. The dependency structure of the parameter prior that was specified by our expert. This dependency structure reflects the relations of the variables according to the subjective knowledge of the expert. The corresponding conditional probability tables specified by our expert are not shown due to lack of space.

receiver operator characteristic curve (AUC) for predicting EPs. Then, these 100 AUCs were averaged, and the combination of priors with the highest average AUC was selected. The ROC curves and the AUCs were estimated (Hanley and McNeil, 1982) and compared (Hanley and McNeil, 1983) using custom scripts written in MATLAB (Version 6.5 Release 13—also see Epstein *et al*. (2002) where the same scripts were applied).

Model training with the optimal combination of priors

This optimal combination of prior information was used to train 100 models on the model-building data set with different initialization parameters. The model with the highest AUC for predicting EPs on the model-building data set was selected.

Model validation on the validation data set

The model that was selected in the previous step was used to predict EPs in the validation data set. The AUC was calculated and represents the performance for predicting EPs of this model on unseen data.

Logistic regression

We compared the developed models by re-training on the model-building data set a previously developed multicategorical logistic regression (Condous *et al*., 2004a) and using this model on the validation data set. M1 consists of a single variable, the HCG ratio, and was built using stepwise logistic regression to predict the probability of failing PUL, IUP and EP simultaneously.

Results

A total of 1003 consecutive women were classified with a PUL between June 2001 and October 2004. Fifty-eight were lost to follow-up, 129 were excluded from the final analysis because of incomplete data and 18 persisting PULs were also excluded from the final analysis. The remaining 856 PULs were used in the final analysis for the model building and testing. The complete data set composed of 460 (53.7%) failing PULs, 330 (38.6%) IUPs and 66 (7.7%) EPs. The data set was split up into a model-building data set and a validation data set as described earlier. The model-building data set had 599 records, of which 44 (7.3%) were EPs. The validation data set contained the remaining 257 samples with 22 (8.6%) EPs. Table Ia and b show descriptive statistics for the model-building and the validation data set.

Table II summarizes the average performance of the four methods of incorporating prior information using randomizations

Table II. The average performance of the four methods of incorporating expert prior information using the model-building data set

Priors	Average AUC (SD)
No priors	0.82 (0.06)
Structure prior	0.81 (0.06)
Parameter prior	0.87 (0.05)
Structure and parameter prior	0.87 (0.05)

This was done by randomizing the model-building data set 100 times for each combination of prior information, in a stratified way, into a set of 70% of the patients used to build the model and a set of 30% to estimate the area under the receiver operator characteristic curve (AUC) for predicting ectopic pregnancies. Then, these 100 AUCs were averaged.

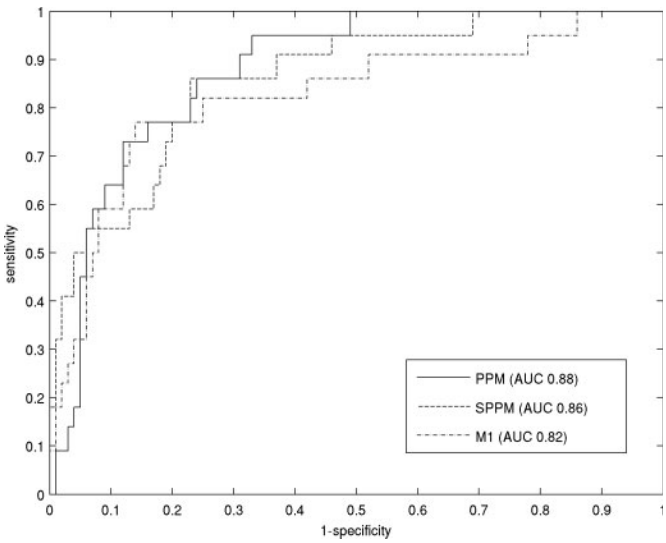

Figure 4. Comparison of logistic regression and Bayesian networks on the prospective data set: the receiver operating characteristic (ROC) curve of the multicategorical logistic regression model (M1) when predicting ectopic pregnancies (EPs) and the ROC curve of parameter prior model (PPM) and structure and parameter prior model (SPPM). The area under the receiver operator characteristic curve (AUC) for these models is given in the figure.

of the model-building data set. The best combinations of priors are (i) using only the parameter prior and (ii) using both the priors (the structure prior and the parameter prior). Their performance was not significantly different (*P*-value = 0.47, Wilcoxon rank sum test), and therefore these two combinations of prior information were selected as optimal.

Next, these optimal combinations of priors (the parameter prior and both the priors) were used to train models on the complete model-building data set. The best model for each combination of priors had the following performance on the complete model-building data set: the parameter prior 0.92 AUC and both priors 0.92 AUC. These models are now referred to as parameter prior model (PPM) and structure and parameter prior model (SPPM), respectively.

Subsequently, PPM and SPPM were tested on the validation data set. Figure 4 shows their ROC curves on this data set. PPM had an AUC of 0.88 and SPPM had an AUC of 0.86. Both ROC curves were not significantly different (Hanley and McNeil, 1983). Figure 5 shows the dependency structures of these two models. We chose the operating point on the ROC curves (resulting from the model-building data set) where the sum of the sensitivity and the specificity was maximal (De Smet *et al*., 2004). The operating point for PPM corresponded to a threshold of 0.13. The probability predicted by PPM is considered an EP above this threshold and a non-EP below this threshold. On the validation set, this gave a sensitivity of 77% and a specificity of 83% for PPM. Next, the likelihood ratios for a positive test result and a negative test result for PPM corresponded to 4.5 (LR+) and 0.28 (LR–), respectively. For SPPM, the operating point was located at a probability of 0.06. A probability predicted by SPPM is considered an EP above this threshold and a non-EP below this threshold. This corresponds to a sensitivity of 77% and a specificity of 80%, and the likelihood ratios for a positive test result and a negative test results correspond to 3.9 (LR+) and 0.29 (LR–), respectively. PPM had a higher AUC and a better specificity than SPPM, and therefore we believe PPM has the potential to be utilized in a clinical setting.

In-depth analysis of the dependency structure of PPM shows that the outcome is separated from the other variables by the HCG ratio, the level of progesterone at 48 h and the number of gestation days (see the theory of d-separation; Pearl, 1988). This means that if the values of these variables are known, then the other variables have no influence on the outcome. This suggests that the HCG ratio, the level of progesterone at 48 h and the number of gestation days are sufficient to predict the presence of an EP. Table IIIa and b summarize the probability of having an EP for the first two states of the gestation day variable (the tables corresponding to the other states of the gestation day variable can be found on the supplementary website http://homes.esat.kuleuven.be/~bioiuser/PUL or this website can be used to select a value for the three important variables and immediately assess the probability of an EP). They visualize and summarize the behaviour of PPM when predicting the probability that a new case is an EP. These tables summarize the probability that a case is an EP for two specific values of the gestation day variable while varying the value of the HCG ratio and the value of progesterone at 48 h. We now highlight two important conclusions that can be drawn from these tables.

First, when the HCG ratio is fixed, there is a similar trend in the tables. When the HCG ratio is below 0.8, then the probability of an EP rises together with rising levels of progesterone at 48 h. When the HCG ratio is equal to or above 1.66, then the opposite trend is seen, the probability drops with rising levels of progesterone at 48 h. Finally, when the HCG ratio is equal to or above 0.8 and lower than 1.66, there is a more complex relationship. The probability reaches a local maximum in the region where the level of progesterone at 48 h is between 10 and 40 nmol/l, but there is a higher probability of diagnosing an EP when the level of progesterone at 48 h is above 80 nmol/l.

Second, the number of gestation days also has a large influence. When this variable is below 35, then the probability of an EP is much higher when the HCG ratio is below 0.8, and the progesterone levels at 48 h are high (Table IIIa), compared to the case when the number of gestation days is above 35 (Table IIIb).

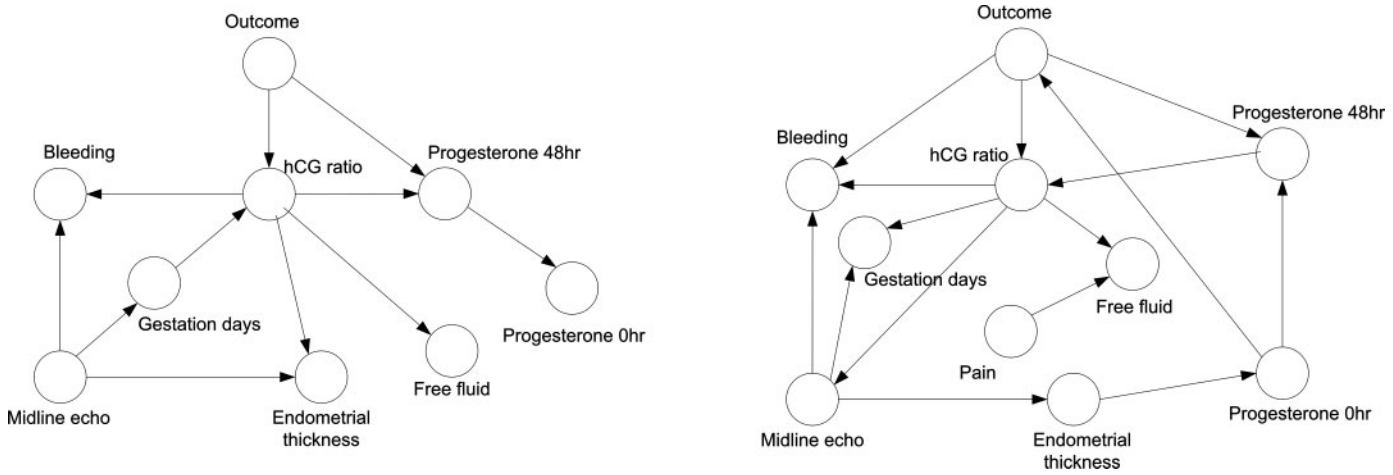

Figure 5. The dependency structure of parameter prior model (PPM) and structure and parameter prior model (SPPM) which where built using the model-building data set. The outcome node represents the outcome variable which can be either ectopic or non-ectopic.

Table III. The probability of diagnosing an ectopic, predicted by PPM, given the number of gestation days is (a) below 35 and (b) equal or above 35 and below 42 while varying the HCG ratio and the progesterone level at 48 h

Progesterone 48 h = Y (nmol/l)	HCG ratio = X		
	$X < 0.8$	$0.8 \le X < 1.66$	$X \ge 1.66$
(a) Number of gestation days below 35			
$Y < 10$	0.05	0.18	0.30
$10 \le Y < 20$	0.43	0.39	0.17
$20 \le Y < 40$	0.56	0.36	0.08
$40 \le Y < 60$	0.82	0.20	0.02
$60 \le Y < 80$	0.82	0.13	0.02
$Y \ge 80$	0.82	0.60	0.02
(b) Number of gestation days equal or above 35 and below 42			
$Y < 10$	0.01	0.14	0.50
$10 \le Y < 20$	0.07	0.31	0.33
$20 \le Y < 40$	0.12	0.29	0.16
$40 \le Y < 60$	0.33	0.15	0.04
$60 \le Y < 80$	0.33	0.10	0.04
$Y \ge 80$	0.33	0.52	0.04

Next, the performance of the logistic regression model M1 when predicting EPs (AUC of 0.82) was compared with PPM and SPPM; however, this difference was not statistically significant (Hanley and McNeil, 1983). Figure 4 shows the ROC curve of M1 when predicting EPs and the ROC curves of PPM and SPPM for comparison.

Discussion

This study demonstrates that Bayesian networks can be used to predict EPs in a PUL population. The model building resulted in two models (PPM and SPPM) that were tested on the validation data set. Figure 4 shows that PPM had the best AUC performance for predicting EPs. The ROC curves of PPM and SPPM also showed that although PPM had a higher AUC, the sensitivity of SPPM remained higher for high specificity compared to PPM. Therefore, SPPM has an advantage. SPPM has higher sensitivity (>40%) at high specificity (>98%) than PPM at this point. This means that it finds more true-positives for fewer false-positives.

The performance of PPM and SPPM (Figure 4) was better than the performance of the logistic regression model M1 when tested on the validation data set; however, this improvement was not significant (Hanley and McNeil, 1983). Assuming that a real difference exists between the true AUC of PPM and M1, the number of patients in the validation set, however, was not sufficient to reach statistical significance. If the AUCs of these models represent the true AUC (i.e. those that would be achieved by infinite populations), one would need a sample size of approximately 350 to detect—with 80% power—the difference between these AUCs as statistically significant. A larger prospective study is thus needed to confirm the difference between these models.

Despite the lack of a significant difference between PPM and M1, the main advantages of Bayesian networks compared to logistic regression are the effect of prior knowledge and the possibility of interpreting the model.

The effect of prior knowledge

The effect of the two expert priors was investigated as a function of the data set size (data not shown). When there are few data, the structure prior helps to learn the model faster. This prior makes it possible to use Bayesian network techniques even when the amount of data is limited. When combined with the parameter prior, the AUC performance for predicting EPs is improved. The parameter prior raises the AUC in the large sample range. Finally, when both priors are used, we have a combined effect: when there are few data, the model benefits from the structure prior and when more data are used, the model benefits from the parameter prior. This shows that subjective prior information from an expert can improve the performance of a model in a profound way both when the size of the data set is small and when the size of the data set is large.

Interpretability of Bayesian networks

The dependency structure of PPM and SPPM is shown in Figure 5. PPM had the highest AUC when tested on the validation data set, and we focus on this model. There is a large difference between the dependency structure of PPM

and the dependency structure specified by our expert as part of the parameter prior (Figure 3). Our expert characterized most of the variables as caused by the outcome variable, whereas in the final model, the influence of most of the variables (excluding the progesterone levels) is mediated by the HCG ratio. This suggests a central role of the HCG variable when predicting EPs.

Furthermore, the absence of the age and the pain variable in PPM suggests that they are not related to any of the other variables. The pain variable could be of low relevance because this is a rather subjective variable. The absence of the age variable was no surprise because this variable was also judged absent by the expert in the structure prior (Figure 2) and in the dependency structure of the parameter prior (Figure 3).

PPM was discussed in detail in the *Results* section. Table IIIa and b together with the tables corresponding to the other states of the gestational age variable (see supplementary website) show the behaviour of PPM. Because only the number of gestation days, the HCG ratio and the progesterone level at 48 h have a direct influence on the outcome (Figure 5), we can evaluate the model at every instantiation of these three variables. This can be done using the tables as charts or using, for example, Microsoft Excel and assess the probability of having an EP, given the number of gestation days, the HCG ratio and the progesterone level at 48 h. These tables can be downloaded from the supplementary website or this website can be used to select a value for the three important variables and immediately assess the probability of an EP. Therefore, these tables are a complete summary of PPM and show the relationship between the outcome and the three important variables: the number of gestation days, the HCG ratio and the progesterone level at 48 h.

The complete PPM model is still interesting because it allows predicting the probability of an EP when there are missing values among the HCG ratio, the level of progesterone at 48 h or the number of gestation days. More generally, PPM allows for the estimation of any missing value in future cases. The full model also shows the relations between the variables and can be updated with future data.

The disadvantage of this methodology is that continuous variables have to be discretized. It is inevitable that some information is lost in the process of discretization. This problem cannot easily be solved because there are often both continuous and discrete variables present when facing clinical decision-support problems. Because mixed Bayesian networks (with both continuous and discrete variables) are mathematically more complex and need approximate methods for classification, we need to discretize the continuous variables or transform the discrete variables to a continuous domain. We chose to do the former because the latter is less intuitive in the case of nominal data (i.e. discrete variables where the data have been classified in qualitative unordered categories, such as the bleeding variable). Moreover, collecting prior information for continuous-valued or mixed Bayesian networks is less straightforward.

Next, when one of the important variables is unknown, the Bayesian networks described in this study need more information. Then, depending on which variable is missing, clinical information (presence of vaginal bleeding and lower abdominal pain), ultrasound information (endometrial thickness and character of the midline echo) or more biochemical information (progesterone levels at 0 h) will be required. Therefore, when one of the important variables is missing, the Bayesian networks might be more costly and time consuming to use in clinical practice, and because, in this case, they partly rely on subjective variables (e.g. pain) and the sonographer's skills, they might be more prone to variation in performance between different centres. However, it is important to stress that, in the presence of missing data for one or more of the three important variables, PPM can still be used to predict the probability of an EP at the cost of having to measure more variables.

We conclude that discrete-valued Bayesian networks can be used to predict the presence of an EP. Furthermore, PPM has extra advantages because it is based on 'more' data through the use of prior information and it incorporates nonlinear relationships between the variables. PPM is highly interpretable and could easily be used. However, prospective interventional multicentre studies are needed to test and compare their performance and cost in different clinical settings.

Acknowledgements

This research is supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen), Research council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, IDO (Genetic networks), several PhD/postdoc and fellow grants, The Flemish Government: FWO: PhD/postdoc grants, G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.05503.06 (VitamineD) and research communities (ICCoS, ANMMM and MLDM); IWT: PhD grants, GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope, Silicos; Belgian Federal Science Policy Office: IUAP P5/22 (Dynamical Systems and Control: Computation, Identification and Modelling, 2002–2006); EU-RTD: FP5-CAGE (Compendium of Arabidopsis Gene Expression); ERNSI: European Research Network on System Identification; Biopattern (FP6-2002-IST 508803); eTUMOUR (FP6-2002-LIFESCIHEALTH 503094) and FP6-MC-EST Bioptrain.

References

Condous G, Okaro E, Khalid A, Zhou Y, Lu C, Van Huffel S, Timmerman D and Bourne T (2002) Role of biochemical and ultrasonographic indices in the management of pregnancies of unknown location. Ultrasound Obstet Gynecol 20 (Suppl. 1),36–37.

Condous G, Okaro E, Khalid A, Timmerman D, Lu C, Zhou Y, Van Huffel S and Bourne T (2004a) The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. Hum Reprod 19,1900–1910.

Condous G, Lu C, Van Huffel S, Timmerman D and Bourne T (2004b) Human chorionic gonadotrophin and progesterone levels for the investigation of pregnancies of unknown location. Int J Gynaecol Obstet 86,351–357.

Condous G, Okaro E, Khalid A, Lu C, Van Huffel S, Timmerman D and Bourne T (2005a) A prospective evaluation of a single-visit strategy to manage pregnancies of unknown location. Hum Reprod 20,1398–1403.

Condous G, Okaro E, Khalid A, Lu C, Van Huffel S, Timmerman D and Bourne T (2005b) The accuracy of transvaginal ultrasonography for the diagnosis of ectopic pregnancy prior to surgery. Hum Reprod 20,1404–1409.

Cooper GF and Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9,309–347.

De Smet F, Moreau Y, Engelen K, Timmerman D, Vergote I, and De Moor B (2004) Balancing false positives and false negatives for the detection of differential expression in malignancies Br J Cancer 91,1160–1165.

Epstein E, Skoog L, Isberg PE, De Smet F, De Moor B, Olofsson PA, Gudmundsson S and Valentin L (2002) An algorithm including results of gray-scale and power Doppler ultrasound examination to predict endometrial malignancy in women with postmenopausal bleeding. Ultrasound Obstet Gynecol 20,370–376.

Hanley J and McNeil B (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143,29–36.

Hanley J and McNeil B (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148,839–843.

Heckerman D, Geiger D and Chickering DM (1994) Learning Bayesian networks: the combination of knowledge and statistical data. In *Proceedings of the 10th Conference Uncertainty in Artificial Intelligence, Morgan.* Kaufmann Publishers, San Francisco, CA, pp. 293–301.

Kadar N, Caldwell BV and Romero R (1981) A method of screening for ectopic pregnancy and its indications. Obstet Gynecol 59,162–166.

Lewis G and Drive J (2004) *Why Mothers Die 2000–2002*. RCOG press, London.

Neapolitan (2005) *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River.

Pearl J (1988) *Probabilistic Reasoning In Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Van Calster B, Condous G, Van Huffel S, Bourne T and Timmerman D (2005) Prediction of pregnancy of unknown location outcome: the difficult yet important case of ectopic pregnancies. In Fonseca JM (ed.) *Proceedings of the Second International Conference on Computational Intelligence in Medicine and Healthcare (CIMED, 2005).* Lisbon, Portugal, pp. 98–105.

Submitted on October 18, 2005; resubmitted on February 8, 2006, February 23, 2006; accepted on March 1, 2006