A Progressive Explanation of Inference in 'Hybrid' Bayesian Networks for Supporting Clinical Decision Making

Evangelia Kyrimi E.KYRIMI@QMUL.AC.UK
William Marsh D.W.R.MARSH@QMUL.AC.UK
School of Electronic Engineering & Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK

Abstract

Many Bayesian networks (BNs) have been developed as decision support tools. However, far fewer have been used in practice. Sometimes it is assumed that an accurate prediction is enough for useful decision support but this neglects the importance of trust: a user who does not trust a tool will not accept its advice. Giving users an explanation of the way a BN reasons may make its predictions easier to trust. In this study, we propose a progressive explanation of inference that can be applied to any hybrid BN. The key questions that we answer are: which important evidence supports or contradicts the prediction and through which intermediate variables does the evidence flow. The explanation is illustrated using different scenarios in a BN designed for medical decision support.

Keywords: Bayesian networks; explanation of reasoning; decision making.

1. Introduction

Many predictive models have been developed in medicine as decision tools (DTs) (Lucas and Abu-Hanna, 1999) but few of them have been used in clinical practice to improve decision making (Wyatt and Altman, 1995; Lucas, 2001). It seems that there is a gap between accurate predictions and a useful DT (Toll et al., 2008; Moons et al., 2009). A contributor to this gap is the perceived trustworthiness of the model, as clinicians will not act on a prediction that they do not trust (Wyatt and Altman, 1995; Moons et al., 2009). The lack of trust may be due to the difficulty of understanding how a prediction is inferred from the given data. As Aristotle wrote 'we do not have knowledge of a thing until we have grasped its why, that is to say, its explanation.' Hence, explaining a model's reasoning – its inference – could increase trustworthiness.

This study considers BNs: a directed acyclic graphs that show causal or influential relationship between the random variables. The variables can be discrete or continuous – a BN with both is called 'hybrid' – and the relationships are expressed using conditional probabilities. When the states of some variables are known, Bayes' theory can be used to infer the updated probability distributions of the others. However, this inference process is not always easy for a user to follow, especially in large networks (Lacave and Díez, 2002, 2004; Pedersen, 2010). The main contribution of this chapter is a method of explaining inference in BNs, so that a user can understand how a prediction is generated. The method can be used in hybrid networks and requires no user input; we call it 'progressive', as the explanation has several levels of detail. The rest of this paper is organised as follows. Section 2 reviews previous research and Section 3 describes our explanation method. Two real scenarios are used as a case study in Section 4. Conclusions and future work are presented in Section 5.

2. Previous Work on Explanation in BNs

In this section we define the term ‘explanation’ and how it can be applied to reasoning in BNs. We survey existing work briefly and explain its limitations.

2.1 What Needs to be Explained?

An explanation is a process of understanding a statement by providing causal connections to known facts (Mayes, 2010). However, different statements require different explanations. Lacave and Díez (2002) reviewed the explanation methods in BNs, distinguishing the following different focuses of an explanation:

- Explanation of the model: we want to explain how the structure and parameters of the model relate to knowledge about the domain.
- Explanation of the evidence: we want to explain the evidence variables by determining the most likely values of the unobserved variables
- Explanation of reasoning: we want to explain how the evidence leads to a prediction for one or more unobserved variables.

All the above types of explanation could be useful to clinicians. However, the aim of this study is to increase clinicians’ trust in the prediction, for which we need to explain reasoning. The form of such an explanation can be illustrated by the following scenario:

> *An emergency doctor has a DT that predicts the likelihood of coagulopathy*[1] *in traumatically injured patients. He enters the evidence and the model predicts that the patient is 8.7 times more likely to become coagulopathic than an average trauma patient. When asked to explain, the system informs him that despite the positive effects of the absence of a long bone and pelvic fracture and the negative fast scan*[2]*, the likelihood of coagulopathy increased because of the thoracic fracture, the high energy of the injury, a base excess*[3] *of -14, a GCS*[4] *of 4 and the administration of more than 500ml of fluids. In complicated cases, just explaining the significant positive and negative causes may not be sufficient: the system can further explain that the evidence affected the prediction of coagulopathy through the unobserved variables tissue injury and tissue perfusion.*

This example shows the basic components of an explanation. First, the explanation has a target, here ‘coagulopathy’. Then the most significant evidence variables that support or contradict the prediction are presented. For more details, the explanation introduces some unobserved intermediate variable through which the information flows and describes how they are affected by the evidence.

2.2 A Review of Existing Methods of Explanation of Reasoning

Several methods of explaining the reasoning in a BN have been proposed. Common elements are i) how to measure the impact of the evidence variables on the target and ii) determine which need to be

[1] Coagulopathy is a bleeding disorder.

[2] Fast scan uses ultrasound to check for internal bleeding.

[3] Using base excess we check for respiratory problems.

[4] GCS assesses the consciousness.

included in the explanation, iii) how to distinguish between supporting and conflicting evidence and finally iv) how to explain the flow of information from evidence variables to the target, described as 'chains of reasoning'. We review the different ways these parts of the explanation problem have been constructed. Other methods (Yap et al., 2007; Sjoerd T. Timmer et al., 2015; Vlek et al., 2016), which lack these parts, are not reviewed here.

2.2.1 THE IMPACT OF EVIDENCE

Not all the evidence has equal impact on the target variable. Measuring the impact involves assessing the change in the probability distribution of the target produced by the evidence; there are different distributions that can be compared and different measures to do that. INSITE uses the KL divergence between the posterior of the target with all the evidence and with each evidence variable (one-way analysis) or a subset of evidence variables (multi-way analysis) removed (Suermondt, 1992). Exact multiway analysis for the best subset of evidence is time consuming as it is exponential to the number of evidence variables. Chajewska and Draper (1998) address the computational complexity with more flexible requirements either for the size of the explanation set or the significance of the impact that each evidence variable has on the target (Chajewska and Draper, 1998). They also point out that the prior probability of the target needs to be considered. BANTER measures the difference between the prior and the posterior of the target for each evidence variable on its own (Haddawy et al., 1997). However, this simplification can be misleading as the effect of the two variables together may be quite different from either of them separately. Madigan et al. (1997) assess the impact using Good's weights of evidence (Good, 1977), evaluated incrementally as the user instantiates each evidence variable; a binary target is assumed and the calculated weights depend on the order in which the evidence is entered (Madigan et al., 1997).

2.2.2 SETTING A THRESHOLD FOR SIGNIFICANT EVIDENCE

The explanation should only include the evidence variables with greater impact. Many different ways have been proposed to find an appropriate impact threshold. A simple approach, proposed by Chajewska and Draper (1998), is for the end user to chose a threshold. However, even if an end user has the domain knowledge needed, it is hard for them to express this in terms of a range of the distance measure. Alternatively, a fixed threshold is chosen by the model builder (Haddawy et al., 1997) or the impact of all the evidence variables is presented, from the largest to the smallest, without a threshold (Sutovský and Cooper, 2008). However, this can make the explanation over complex when there are many evidence variables. Suermondt's INSITE method also proposes an indirect way for the user to choose a threshold. Instead of choosing an appropriate threshold for the distance measure, the user specifies an 'indifference' range for the posterior of the target; changes outside this range are significant and the corresponding threshold can be calculated. This approach combines the users' domain knowledge, given as the range of indifference on the probability, and the characteristics of the distance measure. However, this range may need to be changed for each query and it is still not easy for an end user to do this, especially when the target variable is continuous or the decision tool is being used under time pressure.

2.2.3 SUPPORTING AND CONFLICTING EVIDENCE

We also want to know whether each evidence variable supports or conflicts with the overall change predicted by the model. Suermondt's INSITE introduced the idea of conflict analysis in an explanation, looking at whether removing an evidence variable shifts the posterior of the target in the same direction as the change from the posterior with all the evidence to the prior, with no evidence. However, this analysis is limited to binary variables. For non binary variables mixed effects can occur, where the change for some states supports and for other states conflicts with the overall change. Haddawy's BANTER does not distinguish between the different effects of different evidence. Madigan's use of the weight of evidence distinguishes between positive and negative effects, but it may depend on the order in which evidence is entered.

2.2.4 CHAINS OF REASONING

Evidence variables may be connected to the target by other variables in a 'chain of reasoning'. Choosing which of these variables to include in the explanation is difficult as there can be many such chains. Suermondt's INSITE tool generates a set of directed chains from each significant evidence variable to the target, and, by screening the effect the evidence has on each variable in each chains, it eliminates those chains which block the transmission of evidence. An additional screening is performed by removing arcs that link chains. Haddawy's BANTER selects the chains with the highest strengths and the minimum length (among chains with the same strength) by measuring the impact of every variable in the chain. The strength of the chains is given by the minimum impact of any of the variables in the chain. (Madigan et al., 1997) screen the evidence chains by looking at the weight of evidence of every variable in a chain of reasoning. The weight of evidence for each variable relates to the ratio between the weights of the incoming and outgoing evidence. However, they only consider networks with a tree form, which have only a single path from an evidence variable to the target. Leersum (2015) tries to find a non-empty set of intermediate variables that summarizes all the information between the evidence and the target. He looks at the weight of the edges using a Maximum-flow-minimum-cut theorem and then considers only the variables that are connected with the edges of the minimum cut, which is the minimum set of edges that makes the graph disconnected.

2.3 Improving the Existing Methods

The existing methods have several limitations; we aim to make the following improvements:

- *Support hybrid networks* so that an explanation, including the conflict analysis, can be generated for BNs with both discrete and continuous variables.

- *Reduce complexity* by avoiding exhaustive searches for assessing the impact of evidence and for analysing chains of reasoning, showing that an adequate explanation can be generated that could be used in time-critical applications.

- *Automate significance* so that no user input is needed to determine the threshold for significant evidence.

3. Generating an Explanation in Stages

This section presents an algorithm for a progressive explanation of BN inference.

3.1 Overview

Each explanation has a target (T) and a set of significant evidence (E_{sig}), which is a subset of the evidence (E). The variables that are used in the explanation are called the explanatory variables X. The set X consists of E_{sig} and a set of intermediate variables (X_I) that are unobserved (i.e. not evidence variables) and though which information flows from E_{sig} to T. The different sets of variables are shown in Figure 1.

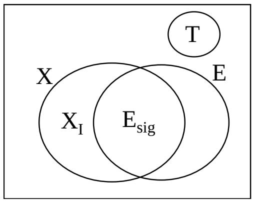

Figure 1: Variables in the explanation of reasoning

The explanation has three levels of increasing detail:

1. The first level lists the significant evidence variables E_{sig}, ordered by their impact on T, showing whether each supports or conflicts with the effect of the combined evidence. This is described in Section 3.2.

2. The second level (Section 3.3) identifies the intermediate variables X_I through which the information from E_{sig} to T flows and it shows how the evidence has changed the probability distribution of X_I.

3. The final level of detail (Section 3.4) describes the effect of each E_{sig} on each of the intermediate variables X_I, showing whether it supports or conflicts with the combined effect.

3.2 Level 1: Significant Evidence Variables

Determining which evidence variables have a significant impact on the target requires first a measure of impact, then a threshold and finally an analysis of whether each item of evidence supports or conflicts with the overall change.

3.2.1 IMPACT OF THE EVIDENCE ON THE TARGET

Following INSITE, the impact of an evidence variable e relates to the distance between the posterior probability with all the evidence E and the marginal posterior probability with e excluded from the evidence.

$$\mathrm{Im}_E(e) \triangleq D_{KL}(P(T|E)||P(T|E\backslash e)) \tag{1}$$

The difference is measured using the well known KL divergence (D_{KL}) . This distance is easy to compute for both discrete and continuous variables. It is u-shaped, giving a greater penalty to the distance from 0.9 to 0.91 than from 0.5 to 0.51. This is appropriate since a probability near either 0 or 1 represents near certainty. The KL divergence is not a true distance, as it is not symmetric, but this is not a disadvantage for our purpose as one of the two distributions – the posterior probability distribution with all the evidence – is fixed. We note that if an evidence variable e is d-separated from T, removing the variable from the set of evidence does not change the probability distribution of T ($\text{Im}_E(e) = 0$). As a consequence, all the variables in E d-separated from T are excluded from the set of E_{sig}.

3.2.2 THRESHOLD OF SIGNIFICANCE

We describe how to extend INSITE's indirect method of specifying a threshold. The threshold θ is the minimum impact, so that:

$$e \in E_{sig} \text{ iff } \text{Im}_E(e) \geq \theta \tag{2}$$

However, rather than giving θ directly, it is defined indirectly using a percentage of indifference α, where $0 \leq \alpha \leq 1$. First, a hypothetical posterior probability distribution G is defined. The distance from the posterior $P(T|E)$ to G is proportional of the distance from $P(T|E)$ to the prior $P(T)$, where G lies in the direction of change (see Figure 2). Finally, θ is defined using the D_{KL} between $P(T|E)$ and the hypothetical posterior G.

$$G \triangleq P(T|E) - \alpha(P(T|E) - P(T)) \tag{3}$$

$$\theta \triangleq D_{KL}(P(T|E)||G) \tag{4}$$

A decreasing list of indifference percentages $\alpha : \{\alpha_1 \ldots \alpha_n\}$ is utilized. Each is used in turn to determine a threshold, continuing until at least half of the evidence variables E are included in E_{sig}.

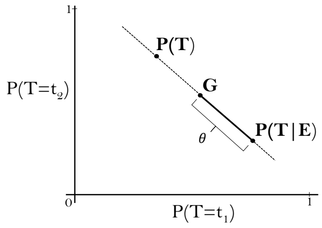

Figure 2: Threshold of significance for a binary target T (based on Suermondt (1992))

3.2.3 CONFLICT ANALYSIS

Having identified the set of significant evidence variables E_{sig}, we next examine whether each evidence variable works in the same way in creating the overall change in T. This is termed 'conflict analysis' and we extend INSITE's method to work for variables with more than two states. We compare (i) the direction of the change and (ii) the impact when each evidence variable is removed with the impact of all the evidence. The direction of change can be assessed using the relative risk $\text{RR}_E(t, e)$ for one state t of the target T and one member e of E_{sig}.

$$\text{RR}_E(t, e) \triangleq P(t|E)/P(t|E \backslash e) \tag{5}$$

For a particular state t, the relative risk of an evidence variable is compared to the relative risk of all the evidence $\text{RR}_E(t, E)$. If a particular evidence variable has the same effect as the overall evidence, we expect each state to be affected in the same way. This gives the following definitions for the consistency with respect to the direction of change:

$$d_{\text{cons}}(e, t) \triangleq \text{RR}(t, e) > 1 \Leftrightarrow \text{RR}(t, E) > 1 \tag{6}$$

$$d_{\text{conf}}(e,t) \triangleq \text{RR}(t,e) > 1 \Leftrightarrow \neg(\text{RR}(t,E) > 1) \tag{7}$$

$$D_{\text{consistent}}(e) \triangleq \forall t. d_{\text{cons}}(e,t) \tag{8}$$

$$D_{\text{conflicting}}(e) \triangleq \forall t. d_{\text{conf}}(e,t) \tag{9}$$

$$D_{\text{mixed}}(e) \triangleq \neg D_{\text{consistent}}(e) \wedge \neg D_{\text{conflicting}}(e) \tag{10}$$

The magnitude of the impact also needs to be considered. If all the evidence is working together, and the direction is consistent, the impact when one variable is unobserved is expected to be less than the impact when all the evidence variables are unobserved ($\text{Im}_E(E)$) i.e. that: $\text{Im}_E(e) \leq \text{Im}_E(E)$.However, it is also possible that removing the evidence e can lead to a greater impact than $\text{Im}_E(E)$, even though the direction is consistent. This suggest that e 'dominates' the remaining evidence. What happens when mixed effects are present? Imagine that we have the target B with three states b_1, b_2, b_3. Three probability distributions are shown in Figure 3, on the left for all the evidence $P(B|E)$ and the right the prior $P(B)$, separated by the distribution with evidence variable e unobserved. For state b_1 the probability is consistently decreasing: $P(b_1|E) > P(b_1|E\backslash e) > P(b_1)$ and for b_3 it is consistently increasing, but b_2 changes first one way then the other. This is the mixed direction.

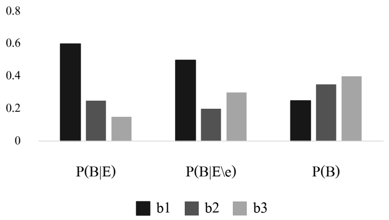

Figure 3: Example of Mixed Effects

To assess which effect is more important, we can calculate the KL divergence over a subset S of the states of the target variable T:

$$D_{KL}(P||Q)_S \triangleq \sum_{i \in S} P_i \log \left(\frac{P_i}{Q_i} \right)$$

We use this to generalise the impact for a subset of states $\text{Im}_E(e)_S$ and test whether the consistent or conflicting states have a greater impact. Table 1 summarise the conflict categories.

Conflict Category	Direction	Impact
Dominant	$D_{consistent}$	$\text{Im}_E(e) > \text{Im}_E(E)$
Consistent	$D_{consistent}$	$\text{Im}_E(e) \leq \text{Im}_E(E)$
Conflicting	$D_{conflicting}$	n/a
Mixed consistent	D_{mixed}	$\text{Im}_E(e)_t \mid t \in d_{\text{cons}}(e,t) > \text{Im}_E(e)_t \mid t \in d_{\text{conf}}(e,t)$
Mixed conflicting	D_{mixed}	$\text{Im}_E(e)_t \mid t \in d_{\text{cons}}(e,t) \leq \text{Im}_E(e)_t \mid t \in d_{\text{conf}}(e,t)$

Table 1: Summary of the Conflict Analysis Categories

3.3 Level 2: Flow of Information

The second level of the explanation uses a simple approach to present the flow of reasoning. First a set of intermediate variables (X_I) is determined: these should be i) unobserved and ii) part of the flow of reasoning from E_{sig} to T. The Markov blanket variables (MB) of T are chosen as the potential set of X_I, excluding those that are observed or are not on a d-connected path from E_{sig} to T, given the evidence variables E. In a BN, the MB of a variable is its parents, children and children's other parents (see Figure 4).

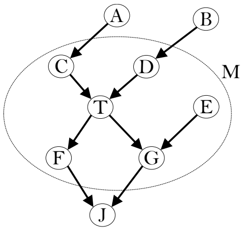

Figure 4: The Markov Blanket (MB) of the Target T is the set $\{C, D, E, F, G\}$

The MB of a variable contains all the variables that shield it from the rest of the network. In the second level of explanation, the change in the uncertainty of each X_I is also shown. If the set X_I is empty (e.g. the whole MB is observed), the explanation stops at the first level.

3.4 Level 3: Effect of Evidence on the Intermediate Variables

The final part of the explanation repeats some parts of the analysis of level 1 on the intermediate variables of level 2. For simplicity and consistency, we do not reassess the set of E_{sig} for each intermediate variable. Instead, for each variable in X_I, the steps are i) determine the subset of E_{sig} that are d-connected to each X_I and ii) carry out the conflict analysis of Section 3.2.3.

4. Case Study

In this section a realistic case study is presented. The focus is not the understanding of the medical disease or how the model was developed; instead, our aim is to show the verbal output of our explanation in the specific case study.

4.1 Detecting Coagulopathy

We show two scenarios with an explanation presented as text. The case study BN was built to predict acute traumatic coagulopathy (COAGUL) in the first 10 minutes of hospital care (Yet et al., 2014). All the variables that may be observed are shown in purple. The target variable, *COAGUL*, is shown in red. There are 11 variables in the Markov blanket of the target. However, the variables *ROTEMA30* to *APTTr* (see top right) are not part in the flow of reasoning and *Death* is also not part of any d-connected path between the target and any of the evidence variables; this results in two intermediate variables: *ISS* (tissue injury) and *PERFUSION* (oxygen in the limbs).

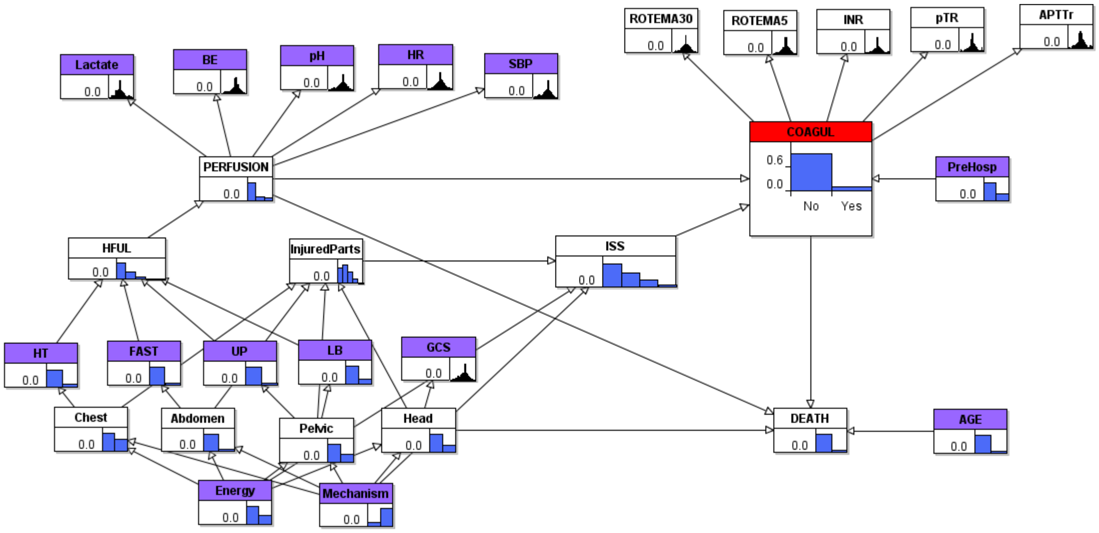

Figure 5: The ATC BN Model

4.2 Scenario 1: Increased Risk with Dominant and Conflicting Evidence

This scenario shows how we explain dominant and conflicting evidence. Before the explanation is presented, the prediction of Coagulopathy is shown, using the absolute and relative risk.

> *The likelihood of 'Coagulopathy' is 11%. This patient has a 14% increase in risk of becoming Coagulopathic, compared to an average trauma call patient.*

Table 2 shows the complete explanation: evidence is described as 'supporting' or 'not supporting'. Supporting evidence has the same effect as the overall evidence on the posterior probability of T. These are the consistent and dominant (distinguished by the phrase 'Very Important') variables of the conflict analysis of Section 3.2.3. Non-supporting variables are those classed as 'conflicting' in Section 3.2.3. In this scenario, where the patient has an increased risk of becoming coagulopathic, the supporting evidence increases the risk, while the non-supporting evidence decreases it. In scenarios, such as scenario 2, where the patient has a decreased risk of becoming coagulopathic, it is the other way round. We therefore use the terms 'support' and 'do not support', instead of 'positive' and 'negative', since they apply in both cases. Table 2 also shows levels 2 and 3 of the explanation: these would be available if a user required more details. Level 2 shows how the intermediate variables X_I have been updated by the evidence: the relative risk of the state with the highest absolute probability is presented. Level 3 shows whether the significant evidence supports the intermediate variables: since 'Tissue Perfusion' and 'Tissue Injury' have more than two states, mixed effects occur and the terms 'partially support' and 'partially do not support' are used for the two categories of mixed effects.

4.3 Scenario 2: Reduced Risk with No Conflicting Evidence

Scenario 2 is a simple case, with no conflicting evidence. The prediction is:

> *The likelihood of 'Coagulopathy' is 0.17%. This patient has a 98% decrease in risk of becoming Coagulopathic, compared to an average trauma call patient.*

Although no explanation may be needed in this situation, the algorithm is able to produce the concise explanation shown in Table 3.

Level 1	
The percentage of change in the uncertainty of Coagulopathy between this patient and an average trauma call patient that is considered insignificant is 50%.	
What are the factors that support the above prediction of 'Coagulopathy'? Factors that support the above prediction of 'Coagulopathy' (strongest to least): • *Pre-hospital fluids $\geq$ 500mls (Very important)* • *GCS = 5 (Very important)* • *Haemothorax = Yes (Very important)* • *Energy of injury = High*	
What are the factors that do not support the above prediction of 'Coagulopathy'? Factors that do not support the above prediction of 'Coagulopathy' (strongest to least): • *Systolic Blood Pressure = 168* • *Long Bone fracture = No* • *Lactate = 0.9*	
Level 2	
How does the model utilize the above factors to predict 'Coagulopathy'? As the immediate causes of 'Coagulopathy' the model uses: *(1) 'Tissue Perfusion': 26% increase in risk of having a Normal 'Tissue Perfusion', compared to an average trauma call patient.* *(2) 'Tissue Injury': 230% increase in risk of having a Severe 'Tissue Injury', compared to an average trauma call patient.*	
Level 3	
(1) Factors that support the prediction of 'Tissue Perfusion': • *Systolic Blood Pressure = 168* • *Lactate = 0.9* • *Long Bone fracture = No* *Factors that do not support the prediction of 'Tissue Perfusion':* • *Haemothorax = Yes*	*(2) Factors that partially support the prediction of 'Tissue Injury':* • *GCS = 5* • *Haemothorax = Yes* • *Energy of injury = High* • *Long Bone fracture = No*

Table 2: The Three Levels of Explanation for Scenario 1

5. Discussion

We have described an algorithm to generate an explanation in three levels, each adding more detail to the explanation. Since the algorithm avoids exhaustive searches it is suitable for real-time use even in complex BNs; it handles both discrete and continuous variables and requires no user intervention. We are currently running a comparative study with clinical users to see if providing an explanation increases their trust in the predictions of the ATC BN or otherwise changes their use of the model. The next steps are to integrate the explanation into a user interface, which may include a graphical representation alongside or instead of the textual one. This will allow us to trial the use of a decision support system with an explanation in real time and examine how much a decision maker under time pressure makes use of the explanation.

Acknowledgement This work was supported in part by European Research Council Advanced Grant ERC-2013-AdG339182-BAYES KNOWLEDGE

Level 1	
The percentage of change in the uncertainty of Coagulopathy between this patient and an average trauma call patient that is considered insignificant is 0.1%.	
What are the factors that support the above prediction of 'Coagulopathy'? Factors that support the above prediction of 'Coagulopathy' (strongest to least): • *Energy of injury = Low* • *Mechanism of injury = Penetrating* • *Fast scan = Negative* • *Haemothorax = No* • *Long Bone fracture = No* • *GCS = 15* • *Pre-hospital fluids < 500mls* • *Systolic Blood Pressure = 157* • *Base Excess = -0.6*	
Level 2	
How does the model utilize the above factors to predict 'Coagulopathy'? As the immediate causes of 'Coagulopathy' the model uses: *(1) 'Tissue Perfusion': 32% increase in risk of having a Normal 'Tissue Perfusion', compared to an average trauma call patient.* *(2) 'Tissue Injury': 78% increase in risk of having a Mild 'Tissue Injury', compared to an average trauma call patient.*	
Level 3	
(1) Factors that support the prediction of 'Tissue Perfusion': • *Systolic Blood Pressure = 157* • *Haemothorax = No* • *Fast scan = Negative* • *Long Bone fracture = No*	*(2) Factors that support the prediction of 'Tissue Injury':* • *Energy of injury = Low* • *Mechanism of injury = Penetrating* • *Fast scan = Negative* • *Haemothorax = No* • *Long Bone fracture = No* • *GCS = 15*

Table 3: The Three Levels of Explanation for Scenario 2

References

U. Chajewska and D. L. Draper. Explaining predictions in Bayesian networks and influence diagrams. In *AAAI Spring Symposium series: Interactive and Mixed-Initiative Decision-Theoretic Systems*, pages 23–31, 1998.

I. Good. Explicativity: a mathematical theory of explanation with statistical applications. In *Royal Society of London A, 354*, pages 303–330, 1977.

P. Haddawy, J. Jacobson, and C. E. Kahn Jr. BANTER: a Bayesian network tutoring shell. *Artificial Intelligence in Medicine*, 10:177–200, 1997.

C. Lacave and F. J. Díez. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, Apr. 2002. ISSN 0269-8889.

C. Lacave and F. J. Díez. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 19(2):133–146, 2004.

J. Leersum. Explaining the reasoning of Bayesian networks, Utrecht University. Master's thesis, 2015.

P. Lucas. Bayesian networks in medicine: a model-based approach to medical decision making. In *EUNITE workshop on Intelligent Systems in patient Care*, pages 73–97, 2001.

P. Lucas and A. Abu-Hanna. Prognostic methods in medicine. *Artificial intelligence in medicine*, 15:105–119, 1999.

D. Madigan, K. Mosurski, and R. G. Almond. Graphical explanation in belief networks. *Journal of Computational and Graphical Statistics*, 6(2):160–181, 1997. ISSN 10618600.

G. R. Mayes. Argument-explanation complementarity and the structure of informal reasoning. *Informal Logic*, 30(1):92–111, 2010. ISSN 08242577.

K. G. M. Moons, D. G. Altman, Y. Vergouwe, and P. Royston. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*, 338:b606, Jan. 2009. ISSN 1756-1833.

K. O. Pedersen. Explanation Methods in Clinical Decision Support, Norwegian University of Science and Technology. Master's thesis, 2010.

Sjoerd T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, and B. Verheij. Explaining Legal Bayesian Networks Using Support Graphs. In *The 13th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 2015. ISBN 9781614996095.

H. J. Suermondt. *Explanation in bayesian belief networks*. PhD thesis, 1992.

P. Sutovský and G. F. Cooper. Hierarchical explanation of inference in Bayesian networks that represent a population of independent agents. In *18th European Conference on Artificial Intelligence*, pages 214–218, 2008. ISBN 9781586038915.

D. B. Toll, K. J. M. Janssen, Y. Vergouwe, and K. G. M. Moons. Validation, updating and impact of clinical prediction rules: a review. *Journal of clinical epidemiology*, 61(11):1085–1094, Nov. 2008. ISSN 1878-5921.

C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij. A method for explaining bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, pages 1–40, 2016. ISSN 1572-8382.

J. C. Wyatt and D. G. Altman. Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ*, 311(7019):1539–1541, Dec. 1995. ISSN 0959-8138.

G.-E. Yap, A.-H. Tan, and H.-H. Pang. Explaining inferences in Bayesian networks. *Applied Intelligence*, 29(3):263–278, Oct. 2007. ISSN 0924-669X.

B. Yet, Z. Perkins, N. Fenton, N. Tai, and W. Marsh. Not just data: a method for improving prediction with knowledge. *Journal of biomedical informatics*, 48:28–37, Apr. 2014. ISSN 1532-0480.