Accepted for publication in *Artificial Intelligence in Medicine*, version 28.02.2016.

Value of Information Analysis for Interventional and Counterfactual Bayesian Networks in Forensic Medical Sciences

Anthony Costa Constantinou[1, 2], Barbaros Yet[2], Norman Fenton[2], Martin Neil[2], and William Marsh[2].

1. Corresponding author. E-mail address: anthony@constantinou.info
2. Risk and Information Management Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, Mile End Campus, Computer Science Building, E1 4NS, London, UK.

THIS IS A PRE-PUBLICATION DRAFT OF THE FOLLOWING CITATION:

Constantinou, A. C., Yet, B., Fenton, N., Neil, M., & Marsh, W. (2016). Value of Information Analysis for Interventional and Counterfactual Bayesian networks in Forensic Medical Sciences. *Artificial Intelligence in Medicine*, 66: 41-52.

DOI: doi:10.1016/j.artmed.2015.09.002
Corresponding author: Dr. Anthony Constantinou, E-mail: anthony@constantinou.info

Abstract:

Objectives: Inspired by real-world examples from the forensic medical sciences domain, we seek to determine whether a decision about an interventional action could be subject to amendments on the basis of some incomplete information within the model, and whether it would be worthwhile for the decision maker to seek further information prior to suggesting a decision.

Method: The method is based on the underlying principle of *Value of Information* to enhance decision analysis in *interventional* and *counterfactual* Bayesian networks.

Results: The method is applied to two real-world Bayesian network models (previously developed for decision support in forensic medical sciences) to examine the average gain in terms of both *Value of Information* (average relative gain ranging from 11.45% and 59.91%) and decision making (potential amendments in decision making ranging from 0% to 86.8%).

Conclusions: We have shown how the method becomes useful for decision makers, not only when decision making is subject to amendments on the basis of some unknown risk factors, but also when it is not. Knowing that a decision outcome is independent of one or more unknown risk factors saves us from the trouble of seeking information about the particular set of risk factors. Further, we have also extended the assessment of this implication to the counterfactual case and demonstrated how answers about interventional actions are expected to change when some unknown factors become known, and how useful this becomes in forensic medical science.

Keywords: Causal inference, Bayesian networks, interventional analysis, counterfactual analysis, value of information, forensic medicine.

1 Introduction

Value of Information (*VoI*) is a technique initially proposed in economics [1] for the purposes of:

1. determining the amount a decision maker would be willing to pay for further information; and

2. prioritising unobserved model factors for acquiring information based on their impact against a desired utility value or probability distribution.

VoI analysis has subsequently been adopted in a number of domains including finance [2], supply chain management [3], pharmaceuticals [4], and health care [5].

An especially important application domain is medicine. For example, *VoI* has been used:

1. as a decision analytic approach to clinical trial design and research priority-setting, by taking into consideration the costs of sampling, the benefits of the sample information, and the decision rules of the cost-effectiveness analysis [6].

2. to determine optimal sample size for clinical trials as an alternative to the more traditional null hypothesis methods [7, 8, 9, 10];

3. for the development and evaluation of clinical trials [11, 12];

4. to investigate the expected value of partial perfect information, and the research decision it can address in medical decision making [13];

5. as a guide to evaluate decision support for differential diagnosis [14];

6. as a decision analysis technique to identify the most beneficial factors in health economic models [5, 15, 16, 17].

For a comprehensive review of *VoI* analyses related to health risk management see [18].

In this paper we are interested in using *VoI* to determine whether missing information can lead to different interventional actions in decision analysis with

Bayesian networks (BNs). The use of *VoI* for interventions has previously been explored in [19] where *VoI* is used to identify novel actions (a process which the authors call *search for opportunities*) in influence diagrams, in the sense that interventions are identified to improve a desirable utility function. More recently, in [20] *VoI* is also considered as an evaluation method for interventional strategies in epidemiology, under competing models, and to quantify the benefit of adaptive versus static intervention strategies. Our major contribution here is to extend *VoI* for interventional decision analysis to the counterfactual setting. This allows decision makers to compare the observed results of the actual world to those of a hypothetical world; i.e. what would have happened had we proposed treatment (or intervention) *B* instead of treatment *A*. To the best of our knowledge, there have been no previous attempts to incorporate the concept of *VoI* to counterfactual problems with BNs.

Our application of *VoI* is motivated by real-world problems in forensic medical sciences in which BNs were developed for decision making. BNs are based on sound foundations of causality and conditional probability theory. Our objective is to show how *VoI* can be applied to BNs to make them especially suitable for simulating interventions and inferring answers from counterfactual questions.

The paper is structured as follows: section 2 describes the forensic medical science problem motivating this work; section 3 provides the necessary background overview of the methods: *VoI*, BNs, interventional and counterfactual analysis; section 4 demonstrates the modelling process of integrating *VoI* analysis into interventional and counterfactual BN decision analysis models; section 5 demonstrates and discusses the results generated by applying the method to two real-world forensic medical case studies; we provide our concluding remarks in section 6.

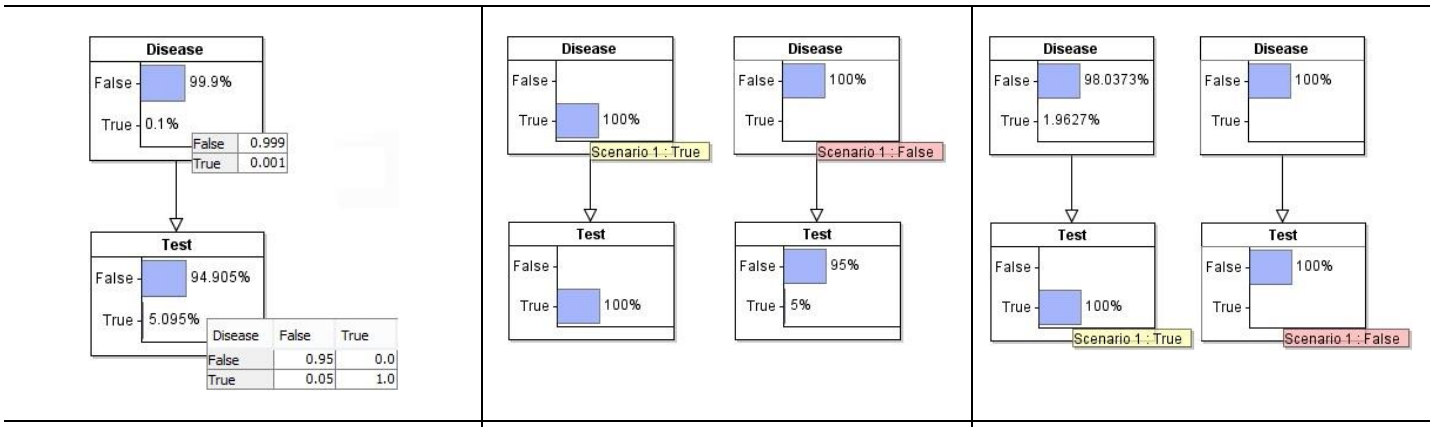
2 Motivation: *The forensic mental health problem*

Forensic medical practitioners and scientists based at the Violence Prevention Research Unit[1] (VPRU); Queen Mary University of London have, for several years, sought improved decision support for determining care and release of people with mental health problems. In particular, they are interested in managing the risk of violent reoffending by releasing such convicted prisoners from prison and discharging such patients from medium secure services [21]. In collaboration with the medical practitioners we have developed two BN models for this purpose – one for prisoners and one for patients [22, 23]. These models delivered significantly improved predictive accuracy with respect to whether a prisoner/patient is determined suitable for release/discharge (hereafter referred to simply as 'release'). The models also provided the additional benefits that causal BN models provide over and above black-box decision models (see Chapters 2 and 3 of [24] for a detailed discussion). However, while, those models were developed for the purpose of simulating interventions (i.e. treatments/therapies) for violence risk management, prior to releasing an individual, they did not consider the possibility that decisions about release could be subject to amendments on the basis of some incomplete information within the model. The BN models were large and complex. Consequently, when assessing an individual for release, information was very often missing for variables that could have been observed[2].

Specifically, a decision maker (such as a probation officer or a clinician) has to determine whether to release a prisoner/patient based on the probability distribution (or the expected value) of the hypothesis variable; i.e. the risk of violence assuming release. Prior to deciding on release, the decision maker has the option to simulate various interventions for the purpose of determining whether an individual's risk of violence can be managed to acceptable levels. Additionally, the decision maker may have the option to gather further information about the individual. While any set of

[1] Formerly known as Forensic Psychiatry Research Unit (FPRU).

[2] Some variables in BNs are supposed to be unobserved. For instance, specific type of latent or uncertain synthetic variables. These also include variables representing symptoms post-treatment, on the basis of some imperfect intervention (see figure 3). In this paper we are only interested in variables with missing information; i.e. those that are not observed, but could have been observed.



1) A BN with no observations; marginal probabilities are displayed. The tables represent the prior probabilities (for the node *Test* this is conditional on *Disease*)

2) A BN with observations about the cause, and posterior probabilities about the effect.

3) A BN with observations about the effect, and posterior probabilities about the cause.

Figure 1. A BN in its simplest form, based on the probability example in [25], demonstrating how prior probability is revised to posterior probability given the specified evidence, from cause to effect and vice versa.

unknown information can still be estimated on the basis of Bayesian inference (via observations provided to other relevant factors within the BN model) it is still possible that knowing (rather than estimating) one or more of these unobserved factors, may lead to amendments in the probation officer's original decision about release.

3 Methods

While a detailed description of the four constituent methods, BNs, *VoI* analysis, interventional and counterfactual analysis is beyond the scope of this paper, this section provides sufficient background to understand the modelling process demonstrated in section 4.

3.1. Bayesian networks (BNs)

BNs, also sometimes known as *belief networks* or *causal probabilistic networks*, are directed acyclic graphical models [26]. They consist of nodes which represent uncertain variables, and arcs which represent causal or influential relationships between the variables. The 'Bayesian' in BNs is due to the use of Bayes' theorem for revising probabilities. Bayes' theorem is a simple equation that specifies how to calculate conditional probabilities:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

where $p(A)$ is the *prior* probability of A and $p(B|A)$ is the likelihood of B given A. The probability $p(A|B)$ is called the *posterior* probability of A. In its prior state all of the variables in a BN are uncertain and assumed to be provisional upon experience/data gained to date. This prior probability is then revised based on new experience/data, to provide the updated *posterior* probability.

Figure 1 presents a very simple BN with just two variables and one dependency. The example is based on a well-known probability problem [25], where a test to detect a disease whose prevalence is 1 in a 1000 has a false-positive rate of 5%. Figure 1.1 presents this problem with both variables being unknown (i.e. the prior marginal probabilities reflecting the average individual). Further, figure 1.2 presents the posterior probabilities for *Test* given the two possible knowns for *Disease*, whereas figure 1.3 presents the posterior probabilities for *Disease* given the two possible knowns for *Test*. While case (2) demonstrates how the cause node affects the probabilities of the effect node, case (3) demonstrates how inference propagates backwards to the cause node having observed the effect, and this is what makes BNs unique for decision analysis. For further reading in BNs see [27, 24].

3.2. Value of Information

VoI analysis was introduced in economics to assess investment decision problems, but is increasingly used in a broader range of applications (see [28]). Figure 2 presents a very simple typical investment decision problem, with BNs. In this example, the question one would expect the *VoI* analysis to answer is "*what is the profit gain of knowing Economic growth prior to making a decision about Investment*".

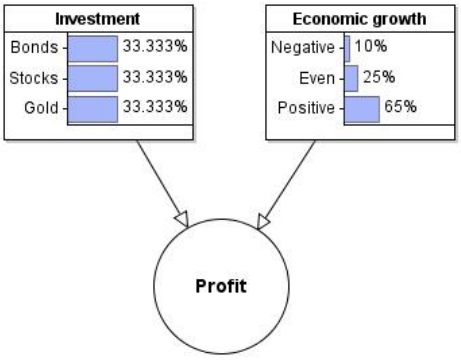


Figure 2. The investment decision problem with BNs.

Table 1 presents the expected monetary value for *Profit (p)*, given *Economic growth (e)* and *Investment (i)*. The table indicates that returns from bonds are independent of fluctuations in economic growth (over some fixed period of time). On the other hand, fluctuations in economic growth are expected to affect the returns from investing in stocks and gold.

Table 1. Table for *Profit p*, where *e* is *Economic growth* and *i* is *Investment*.

e	*Negative*			*Even*			*Positive*		
i	*Bon.*	*Sto.*	*Gol.*	*Bon.*	*Sto.*	*Gol.*	*Bon.*	*Sto.*	*Gol.*
p	£30	-£1000	-£300	£30	£50	£75	£30	£400	£150

If we randomly make an investment decision i, without further information, the expected value (EV) for p is £96.25 (i.e. each action is assigned a prior probability of $\frac{1}{3}$). Given that e is unknown, we would like to make a decision about investment i that maximises profit p. This is defined as the *expected maximum value* (EMV); thus,

$$EMV_p = \max_i \sum_e p_e R_{i,e} = 172.5$$

where $R_{i,e}$ is the payoff table for row index i (which describes the possible actions for the decision maker, i.e. the investment decision) and column index e (which describes the uncertain variable for which the decision maker does not have knowledge, i.e. the unknown *Economic growth*), that has probability p_e of being in state i. In the BN example presented in figure 2, the equivalent required calculation process[3] would be to simply iterate through observations i and pick the observable state that maximises p, while e is uncertain.

However, we want to know the gain when e is known, ***prior*** to making a decision for i. This is defined as the *expected value of perfect information* (EVPI). Specifically, this is EV_p given perfect information (PI) for e (or $EV_p|PI_e$), where for each possible direction of e the investment decision i that maximises p is always selected as shown below:

$$EV_p|PI_e = \sum_e p_e \left(\max_i R_{i,e}\right)$$

Therefore, if we were able to know the direction of economic growth we would have expected to increase our EV for p, on average, from £172.5 to £281.75. Thus, knowing e is worth:

$$EV_p|PI_e - EMV_p = £281.75 - £172.5 = £109.25.$$

Specifically, the maximum value the decision maker should be willing to pay for perfect information is £109.25.

By definition, the *EVPI* represents the potential gain associated with having perfect information on *all* of the unknown model factors. For the decision maker, it is often more useful to assess *VoI* for individual or subsets of unknown factors, rather than assessing all unknown factors collectively. This is defined as the *expected value of partially perfect information* (*EVPPI*); this has the same equation to that of *EVPI*, but for a selected subset of unknown model variables. In the simple example demonstrated above there was only a single unknown factor and so the *EVPPI* of knowing that individual factor is equivalent to the *EVPI*. In this paper, however, we are interested in more complex models with multiple unknown factors and, therefore, we are interested in the *EVPPI*.

3.3. Interventional analysis

Interventional analysis in BNs enables decision makers to prioritise interventions based on evidence [29, 30]. An intervention is an action that can be performed to manipulate the effect of some desirable future outcome. In medical decision analysis, an intervention is typically represented by some treatment, which can affect a patient's health outcome. These are typically described as *imperfect* interventions; implying that the intervention induces a distribution over outcome states, rather than a specific state; i.e. *perfect* interventions [31, 32]. As a result, the effectiveness of an imperfect intervention is expected to be dependent on some other factors. In medicine, other such factors can be *responsiveness* and *motivation* for treatment [22, 23]. We will call these *interventional factors*.

[3] AgenaRisk, which is a BN tool (Agena, 2015), allows the user to run the three different scenarios in a single model. AgenaRisk also allows continuous variables to be incorporated into the model (i.e. such as the *Profit* node in figure 2) without any constraint for static discretisation and with the ability to define any statistical distribution. This is achieved by the use of the dynamic discretisation algorithm (Neil et al., 2010) which uses entropy errors as the basis for approximation.

Figure 3 demonstrates an example of how an observational BN model is expected to change when it is used for imperfect interventional analysis. The top model represents the observational phase and assumes:

1. delusions and anxiety may cause anger;

2. anger and lack of self-control may cause uncontrolled aggression;

3. treatment for anger, which in this case represents the doctor's historical frequency in prescribing treatments, depends on symptoms of anger and/or uncontrolled aggression;

4. motivation to attend treatment and responsiveness to treatment typically depend on the type of treatment (there could be multiple mutual exclusive treatments for anger).

The bottom model represents the interventional phase. Any parent links from causes entering intervention (such as *Treatment for anger*) are expected to be removed. This is because we do not want to infer posterior probabilities for causes when indicating an intervention, which has to be either *true* of *false*, as it happens in the observational phase where the model proceeds to explain the observation for treatment.

In the example of figure 3, the intervention imperfectly manipulates *Anger* and this assumes that some other interventional factors exist, which determine the effectiveness of the intervention. In our case these factors were (according to the experts) *Motivation to attend treatment* and *Responsiveness to treatment*; implying that the transformation from an observational to an interventional model is not always deterministic, at least in the case of imperfect interventions. As a result, both of these interventional factors influence *Anger post-treatment* in the interventional phase, since the effectiveness of *Treatment for anger* is dependent upon them. The example also demonstrates that it is possible for the intervention to serve as the child node of the *Effect* in observational models, but this link should be reversed (if not removed) in the interventional model. Specifically, in the observational model we would expect evidence of *Anger* to increase the chance for a doctor to propose *Treatment for anger*, whereas in the interventional model we would expect *Treatment for anger* to manipulate symptoms of *Anger*. This does not appear to have been discussed in previous relevant research [30, 31, 32].

We have demonstrated the process of an imperfect intervention. If the intervention perfectly manipulates the effect node then the decision maker may express this directly into the effect node, rather than specifying a new probability function, in which case the effect node is now the variable that should be manipulated independently of its causes. The process of intervening on an event that becomes independent of all its causes is known as *graph surgery* [30].

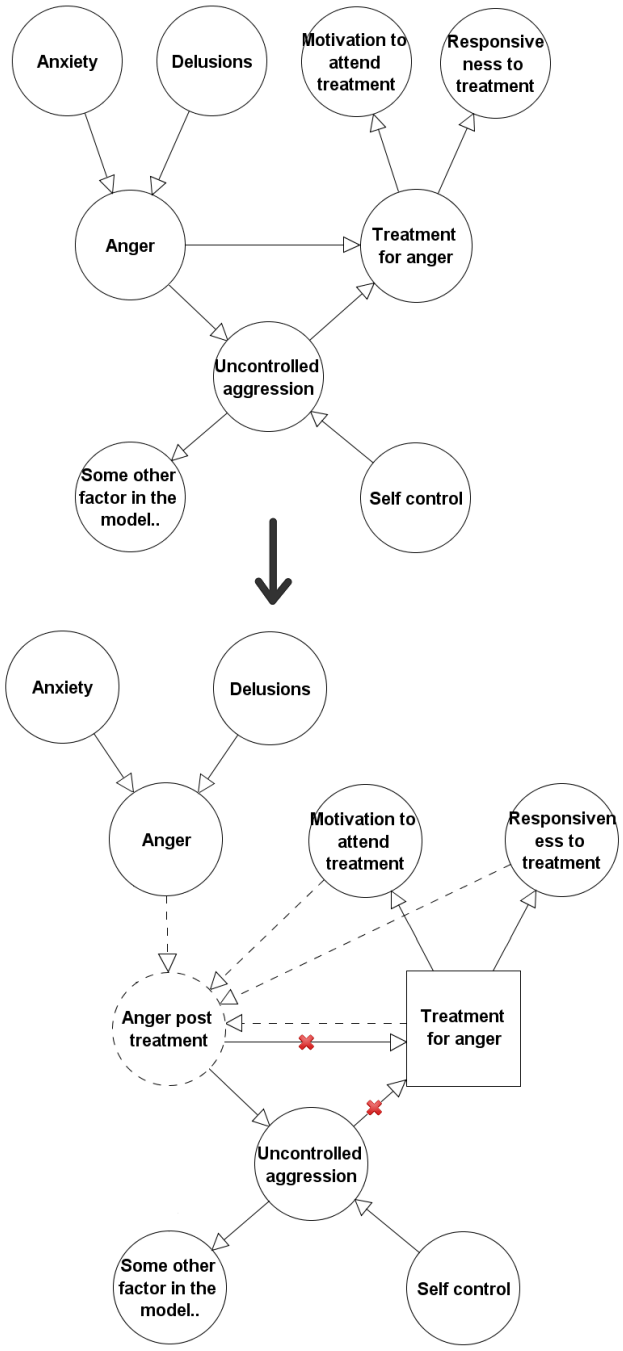


Figure 3. An example of how an observational BN model transforms into an interventional BN model [33]. Dashed nodes[4] and arcs are introduced in the interventional phase.

[4] The node *Anger* serves as the prior for *Anger post-treatment*.

3.4. Counterfactual analysis

Counterfactual analysis in BNs enables decision makers to compare the observed results in the real world to those of a hypothetical world; what actually happened and what would have happened under some different scenario.

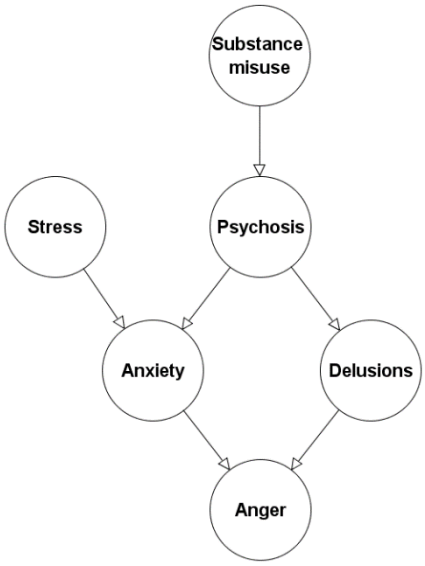


Figure 4. The example BN model prior to performing counterfactual analysis.

Let us consider the BN model presented in figure 4. In this example, we are interested in the outcome of the node *Anger*. Suppose that we observe that *Anger* is *true*, without knowledge of either *Anxiety* or *Delusions*, but with knowledge that *Stress* is also *true*. We want to answer the following question: "*given that Anger and Stress are true, what is the probability that Anger would (still) have been true if we had also known that Anxiety was false?*". To answer this counterfactual question, we make use of the *twin-network* method proposed by Pearl [30]. However, it should also be noted that Dawid [34] proposed a decision-theoretic alternative, which is based on the argument that potential outcomes are inherently metaphysical and that counterfactual models for causal inference can be misleading. These issues relating to counterfactuals and causal inference are discussed further in [35, 36].

Figure 5 demonstrates the application of the twin-network method, where a network representing the actual world is connected to another (twin) network representing the hypothetical world. Both networks share the background variables *Substance misuse* and *Stress*, since those remain invariant under modification [30]. *Anxiety* represents an observation in the actual world, whereas in the hypothetical world we are intervening on *Anxiety'* and hence, we have to follow the process of graph surgery, whereby the variable under *perfect* manipulation becomes independent from its causes. As a result, dashed arcs entering *Anxiety'* are removed. This leaves us with a model in which *Anger* and *Stress* are *true* in the actual world, and *Anxiety'* and *Stress* are respectively *false* and *true* in the hypothetical world. The variable *Anger'* in the hypothetical world provides the answer to the counterfactual question.

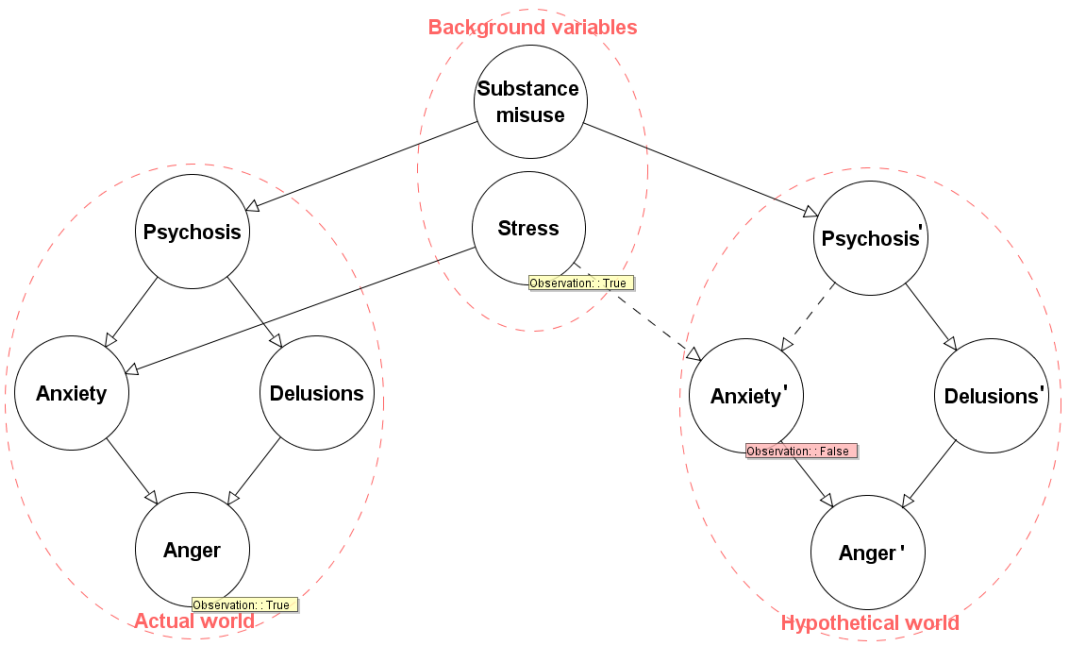


Figure 5. Applying the twin-network modelling technique to the BN model of Figure 4 to answer the counterfactual question about *Anger'*. Dashed arcs entering *Anxiety'* in the hypothetical world are removed since the specified variable is under *perfect* manipulation.

4 The modelling process

In developing interventional and counterfactual BNs we first require the standard observational model that is typically used for prediction. From there, we can construct interventional and counterfactual BNs as described in section 3. The process is presented in figure 6 and illustrates how we may proceed from an observational BN model to:

1. counterfactual analysis without interventions,
2. interventional analysis with interventions, and subsequently to
3. counterfactual analysis that incorporates those interventions.

At each of these stages decisions will be analysed, and *VoI* analysis can be used to determine whether amendments are expected in the decisions under analysis, on the basis of some incomplete information within the model. In this paper, we are interested in modelling stages beyond the observational phase.

To demonstrate the process, the starting observational model is a simplified version of the real-world BN models from the domain of forensic psychiatry (see top part of figure 7) that were described in section 2. These models concern individuals with serious mental health problems who are about to be released. By using the above modelling process we can perform the following decision and risk analysis:

1. *Risk assessment*: The observational BN model is used to assess the risk of violence for the given individual, over a specified time period, in case of release.

2. *Risk management*: An interventional BN model is used to examine whether the risk of violence for the given individual can be managed to acceptable levels. This is achieved by simulating interventional actions to manipulate, directly or indirectly, the estimated risk of violence.

3. *What if risk analysis*: A *counterfactual* BN model is used post-release to study individuals who were violent, and examine whether their risk of violence could have been managed better at the assessment phase.

In [22, 23] only (1) was covered and thus, the new modelling process provides a much enhanced decision-support system. In what follows we focus on (2) and (3) which are respectively demonstrated in subsections 4.1 and 4.2 by incorporating the *VoI* analysis concept.

The decision maker may be interested in whether a decision about an interventional action could be subject to amendments on the basis of reducing model uncertainty. This interest is also extended to the counterfactual case whereby decision analysis based on counterfactual manipulations may also be subject to amendments. *VoI* analysis is used to determine whether there are expectations that would make seeking information about those unknown risk factors worthwhile prior to suggesting any interventional actions.

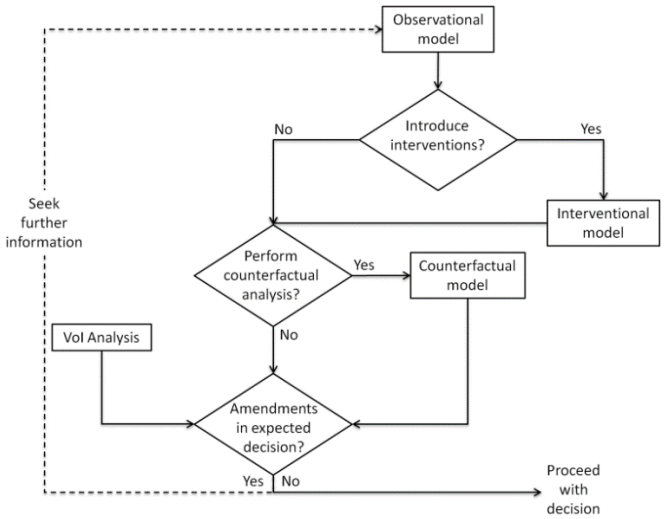


Figure 6. The complete modelling process. The dashed link highlights the stage at which we determine whether further information is required prior to proceeding with the decision suggested by the model.

Figure 7 shows how the simplified forensic medical model is modified from an observational model into its interventional stage; i.e. after *graph surgery* has been performed (as demonstrated in subsection 3.3). We are interested in the interventional model. This simplified version is presented with subjective probabilities for the sake of demonstration. The conditional probability tables (CPTs) for each node[5], of the interventional model, are provided in appendix A.

Specifically, in this simplified version we consider that:

1. A decision to release an individual depends on the individual's risk[6] of violence. We assume: *no release* if this risk is >50%; *release* if <20%; otherwise *release with supervision*. This decision making occurs in the observational phase. In the interventional phase, a

[5] No CPTs are provided for interventional variables since they can either be *true* (perform intervention) or *false* (do not perform intervention).

[6] In this example we simply use the *EV* of the probability distribution for decision making. Other alternatives include the probability distribution itself, or some expected utility derived from the probability distribution.

decision about release is modelled as an intervention that either reduces or eliminates the risk of violence.

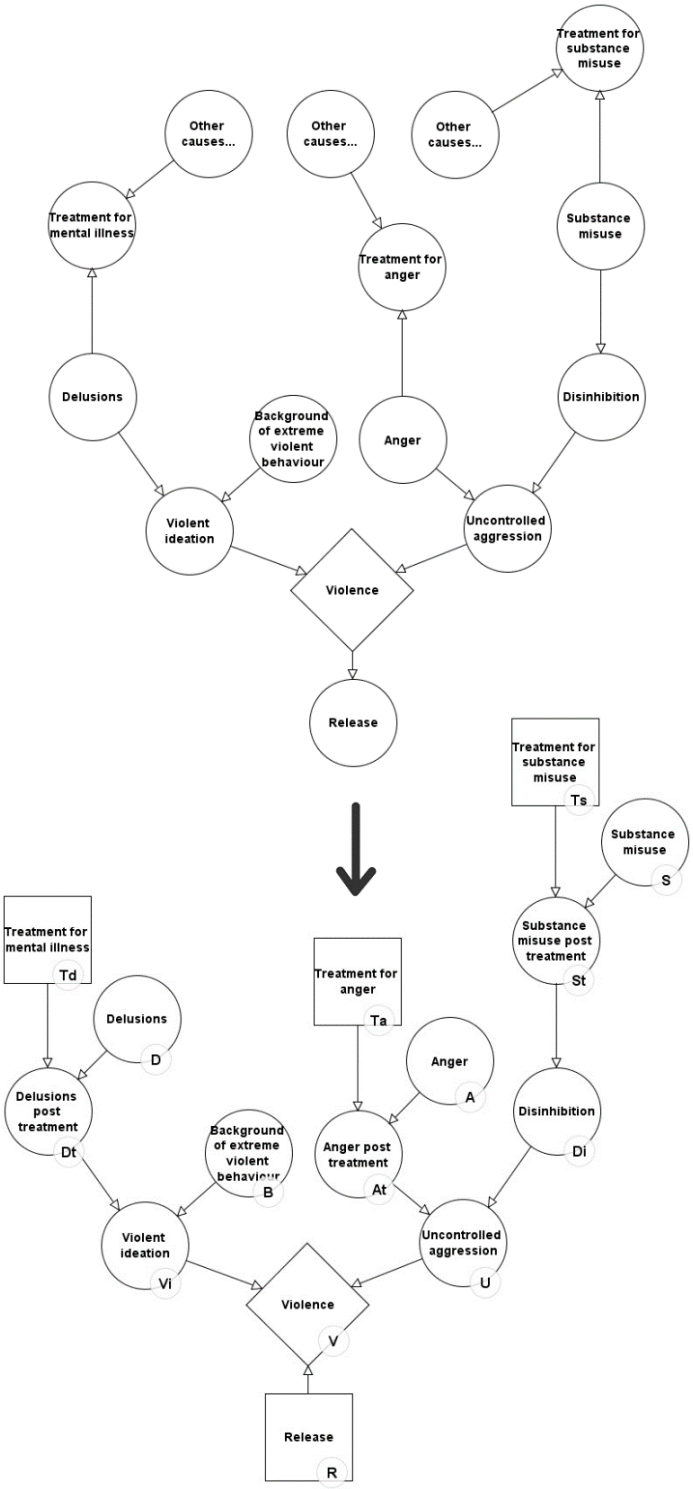


Figure 7. An example of how an observational BN model transforms into an interventional BN model, based on a simplified forensic mental health BN in [22, 23]. Squared nodes indicate interventions, whereas the diamond-shaped node is treated as a probabilistic utility node on which the decision of R is based, in the observational phase, and which either perfectly or imperfectly manipulates V, in the interventional phase.

2. Violence depends on *violent ideation* and *uncontrolled aggression*.

3. Violent ideation depends on *delusions* and *background of extreme violent behaviour*.

4. Uncontrolled aggression depends on *anger* and *disinhibition*, which is caused by *substance misuse*.

5. Delusions, anger and substance misuse can be imperfectly manipulated, in the interventional model, with the respective interventions of *Treatment for mental illness*, *Treatment for anger*, and *Treatment for substance misuse*. In the observational model, treatments are observational variables that simply indicate the probability of a particular treatment to be suggested on the basis of one or more relevant symptoms.

In what follows we make use of the *VoI* abbreviations from subsection 3.2, and the notation provided to the model variables of figure 7. The variable V, which represents the probability an individual is violent post-release, is treated as a utility node on which *VoI* analysis is performed. However, we are not directly interested in the fluctuations of V, as in the standard concept of *VoI*, but rather whether such fluctuations are sufficient for the *expected decision (ED)* for action R to change, on the basis of reducing model uncertainty.

4.1. *VoI for Interventional Bayesian networks*

Interventional analysis is understood to be particularly useful in terms of decision making for release. For example, an individual who is believed to pose a high risk of violence could still be released if the model indicates that his risk of violence can be managed to acceptable levels on the basis of some intervention/s.

Due to the size and complexity of the BN models, interventional analysis is typically performed with some missing information about an individual's risk factors for violence. We would like to know whether the gain from knowing one, or more, of these risk factors is sufficiently strong to amend a desired interventional action.

Suppose that we know that a particular individual under assessment for release a) suffers from delusions, b) suffers from anger problems, and c) has a background of extreme violent behaviour. At this stage, relevant treatments are yet to be suggested; implying that all of the interventions are still set to *false*.

Based on the above set of information, the $EV_V = 0.670$; implying *no release*. The expected minimum[7] value for V (EmV_V) can be calculated by iterating through the

[7] Note that for this example we are interested in minimising the target value (i.e. the risk of violence), rather than maximising profit.

possible states for *Td* and *Ta*, which represent the treatments that could be suggested based on known relevant symptoms, and selecting the combined set of states that minimise *V*; i.e. when both treatments are *true*.

Specifically,

$$EmV_V = \min_{Td,Ta} \sum_S p_S R_{Td,Ta,S} = 0.5034$$

As a result, the model suggests *no release* even when accounting for EmV_V.

However, we have no information about *S*. We would like to know whether it would be sensible not to release the individual without knowing *S*. To find this answer we have to calculate $EV|PI_S$, and in doing so take into account *Ts* as an additional minimiser for *V*. Specifically,

$$EV_V|PI_S = \sum_S p_S \left(\min_{Td,Ta,Ts} R_{Td,Ta,Ts,S} \right)$$

$$EV_V|PI_S = 0.2 \times 0.474 + 0.8 \times 0.4850 = 0.4828$$

As a result, knowing *S* is expected to reduce the *EmV* for *V*, on average, from 0.5034 down to 0.4828. Thus, the gain from knowing *S* is worth:

$$EmV_V - EV_V|PI_S = 0.5034 - 0.4828 = 0.0206$$

but in terms of decision making for *R*, knowing *S* is expected to amend action *R* since $EV_V|PI_S < 0.5$. We will refer to this outcome as the expected decision for *R* (ED_R); i.e. the ED_R when *S* remains unknown is *no release*, whereas when *S* becomes known then the ED_R becomes *release with supervision*. This may sound counterintuitive on the basis that knowledge of *S* might mean knowledge that the individual is a substance misuser. However, the key concept here is that if we had known that the individual was a substance misuser, we would have arranged for a suitable treatment; whereas without knowing *S* it is impossible to arrange such a treatment and thus, we risk not treating the individual in the case where *S* is *true*.

This simple example demonstrates how the concept of *VoI* analysis can be implemented to help the decision maker avoid suggesting potentially erroneous interventional actions on the basis of ignoring some unknown risk factors that could amend a decision for a given interventional action.

4.2. *VoI for Counterfactual Bayesian networks*

Counterfactual analysis can provide clinicians and probation officers who work in these areas the ability to assess a case whereby an act of violence that has already been observed post-release could have been managed better if they had known some further information that was not considered at the assessment phase. This could possibly serve as a lesson learnt for future such cases.

Consider the example of an individual who was violent after release, in which at the assessment phase for release, the probation officer concluded that the individual a) did not have a background of extreme violent behaviour, and b) did not suffer from substance misuse. This led to the decision to release the individual without the need for any sort of treatments, implying that all of the three interventional actions were set to *false*.

After the individual became violent, a new piece of information is observed: that the individual's act of violence was based on cultural, ethnic, and religious incentives. This factor was not considered by the model at the assessment phase. If the probation officer had considered this factor, and had also been aware of the individual's incentives, he could have used the model to analyse the revised risk of violence by also considering relevant interventions for managing the risk of this factor, such as some sort of spiritual care. The relevant counterfactual question in this case is: *Given that the individual was violent post-release, what would be the probability for violence had we also known that he had Cultural, ethnic, and religious incentives for violence and, on that basis, the probation officer had instructed some sort of spiritual care.*

This counterfactual problem now involves external factors that must be taken into consideration. This requires a slightly revised model which incorporates the additional factors. The resulting new/revised CPTs considered for demonstrating this counterfactual case are provided in appendix B.

Our observations in the actual world now consist of both those observed at the assessment phase as well as those observed post-release. These are presented in figure 8, which shows how the model of figure 7 is modified using the twin-network method described in section 3.4. Figure 8 also provides additional notation for the new variables.

From the assessment phase we know:

S = *false*, *Ts* = *false*, *Ta* = *false*, *Td* = *false*, and *B* = *false*.

We also learnt that post-release:

$V = true$ and $C = true$ (the new variable).

On the other hand, the counterfactual question of spiritual care comes into effect only in the hypothetical world. At this stage, there are seven background variables; i.e. the five presented in the middle section of figure 8, plus Td and D (the dashed nodes, Td' and D', come into effect later with a modified version of this example). The $EV_{V'}$= 0.5577 prior to suggesting spiritual care on the basis of C', whereas the

$$EmV_{V'} = \min_{Sc} \sum_{D'} p_{D'} R_{Sc,D'} = 0.5252$$

where Sc minimises V'; i.e. if the probation officer had suggested spiritual care. Therefore, with $EmV_{V'}$ providing the answer to the counterfactual question we can conclude that the particular individual would not have been released had we have known his cultural, ethnic, and religious incentives, even when accounting for spiritual care.

However, this suggestion comes without having information about either delusions or anger. We would like to know whether the suggestion for hypothetical action *no release* is expected to be subject to amendments if we had reduced model uncertainty. We test this in the case of delusions. Hence, the model requires some modification.

Figure 8 indicates how the twin-network model is altered in order to manipulate delusions in the hypothetical world. Specifically, at this stage the dashed nodes Td' and D' are introduced, whereas the previously background variables D and Td become specific to the actual world and thus, the dashed links entering Dt' are removed. We therefore have to calculate $EV|PI_{D'}$ and in doing so, we also have to take into account Td' as an additional minimiser for V'. As a result, in this extended counterfactual case, Dt' is dependent on the new set of variables D' and Td'. The EV for V' when D' is known is

$$EV_{V'}|PI_{D'} = \sum_{D'} p_S \left(\min_{Sc,Td'} R_{Sc,Td,D'} \right)$$

$$EV_{V'}|PI_{D'} = 0.6 \times 0.4656 + 0.4 \times 0.5176 = 0.4864$$

Thus, while the gain from knowing D' is just worth

$$EmV_{V'} - EV_{V'}|PI_S = 0.5252 - 0.4864 = 0.0388$$

the discrepancy is enough to alter the $ED_{R'}$ from *no release* (when D' was unknown) to *release with supervision* (when D' is known).

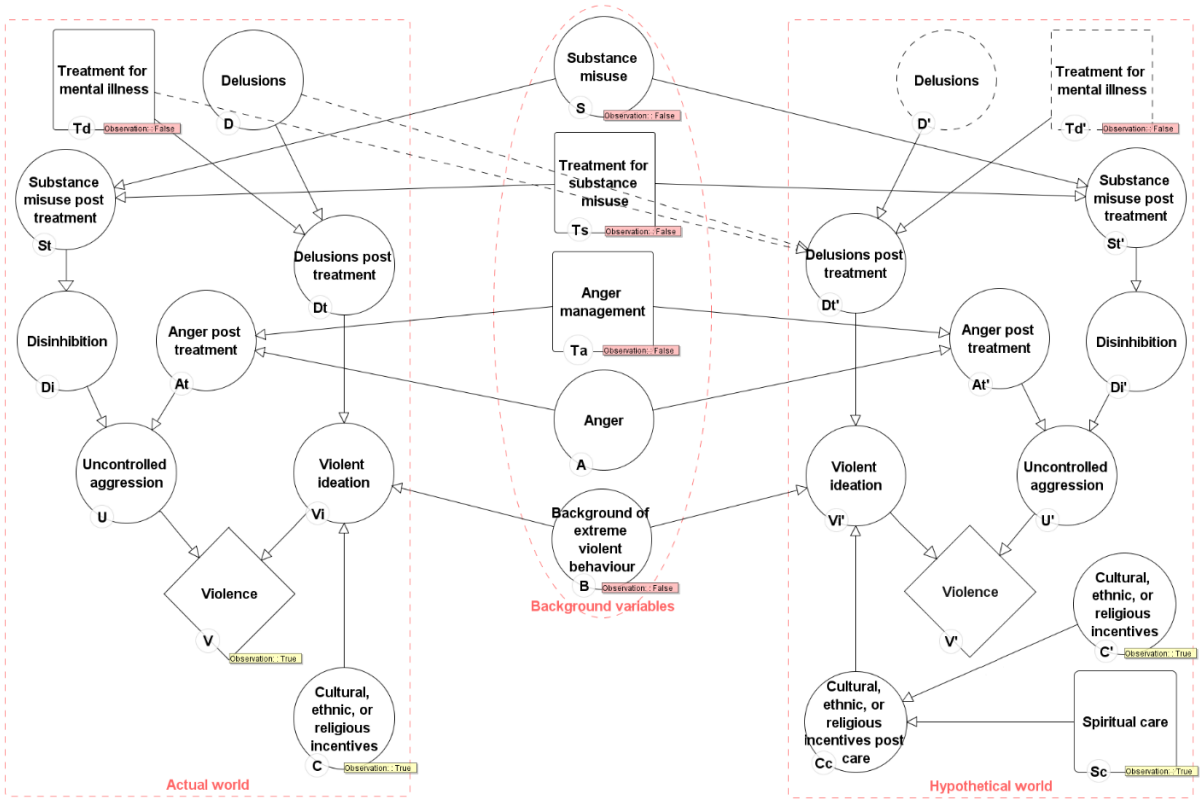


Figure 8. Using the twin-network method to answer the counterfactual question for V'. The counterfactual BN model also indicates the modifications required to accommodate the $EV_{V'}|PI_{D'}$ as discussed in subsection 4.2 (i.e. a new instance of the background variables Td and D is created in the hypothetical world, and dashed links entering Dt' are removed).

5 Real-world case studies and discussion

In the previous section we demonstrated how the concept of *VoI* analysis can be used to examine the implications of model uncertainty on decision analysis for interventional actions, in both interventional and counterfactual BN models. The implications have been demonstrated on the basis of limited fluctuations of the *EV* of the hypothesis variable on which an interventional action of interest is based, but which were sufficient for the *ED* to change. In an effort to keep the examples simple, the *VoI* analysis was restricted to a single unknown risk factor that could only be manipulated by a single intervention.

In this section we assess the implications further by applying the method to the two BNs discussed in section 2, which represent two real-world applications to forensic medical problems, and examine the average gain one would expect to observe in terms of both *VoI* and decision making based on real data. The model presented in [22] is called DSVM-MSS, and the model presented in [23] is called DSVM-P. We will use these two terms to distinguish between the two models.

We have performed six experiments in total; two for DSVM-P, one for the interventional and another for the counterfactual case, and four for DSVM-MSS, two for the interventional and two for the counterfactual case. The experiments are double in the case of DSVM-MSS because it incorporates two variables of interest; *violent convictions* and *general violence*, whereas DSVM-P only assesses *violent convictions*. The experiments are summarised as follows:

a) ***The interventional case*:** This is done by examining the average percentage gain expected by simulating a number of interventions on relevant symptoms that can be manipulated by these interventions, as defined within the models. Experiments 1, 2, and 3 reported in table 2 represent the cases of DSVM-P for violent convictions, and DSVM-MSS for violent convictions and general violence respectively.

b) ***The counterfactual case:*** This is done by examining the average percentage gain expected by simulating a number of interventions on relevant symptoms in the hypothetical world, as defined within the models, and after having observed that an individual had been violent post-release. Therefore, this assessment is restricted to cases for which the individuals have been found to be violent over the follow-up period (an average of 5 years for the DSVM-P study and an average of 1 year for the DSVM-MSS study). Experiments 4, 5, and 6 reported in table 2 represent the cases of DSVM-P for violent convictions, and DSVM-MSS for violent convictions and general violence respectively.

The experiments assume that the variables targeted for intervention (eleven for DSVM-P and five for DSVM-MSS) are unobserved at the observational phase. The experiments also assume that, in the case whereby an intervention manipulates multiple variables, at most one such variable is observed (which is enough to activate the intervention) at random, during both the interventional and counterfactual assessments.

Table 2. Absolute (ABS) and relative (RLT) gain observed, in terms of *p(O)* being *true*, for each of the experiments described in section 5, where *VtI* is the number of variables (i.e. symptoms) targeted for intervention, *I* is the number of available interventions, and *p(O)* is the initial average probability for the outcome of interest (i.e. violent convictions or general violence).

E	Data instances	VtI	I	*p(O)*	ABS Gain	RLT Gain
1	953	11	4	32.94%	-03.77%	-11.45%
2	386	5	3	03.25%	-00.50%	-15.27%
3	386	5	3	12.03%	-05.35%	-44.49%
4	240	11	4	32.65%	-12.48%	-38.60%
5	11	5	3	03.34%	-00.68%	-20.42%
6	44	5	3	14.82%	-08.88%	-59.91%

The results from table 2 show that while DSVM-P incorporates a higher number of variables available to be targeted for intervention, as well as a higher number of interventions, it does not seem to generate a higher relative average gain compared to DSVM-MSS. Essentially, the average gain also depends on the structure of the network, the impact of the interventions as defined within the model, as well as on the application domain.

Figure 9 demonstrates the average gain in terms of *ED* for the average individual and for each of the six experiments. The results are separated into four different threshold levels in determining release based on the risk of violence; i.e. when Θ=0.1 we assume that the individual is suitable for release if his or her risk of violence is lower than 10%. The dashed line indicates the shift towards an increased chance of the average individual determined as being suitable for release. Specifically, and based on various decision thresholds, the potential amendments in decision making ranged between 0% and 18.73% for the interventional case, and 0% to 86.80% for the counterfactual case. Note that the gain is 0% only for the cases whereby 100% of the individuals had already been identified suitable for release, according to the hypothetical threshold levels, prior to examining any potential interventions.

The results show that, while the level of gain depends on the model and outcome under assessment,

the method is generally useful in terms of supporting the decision makers with regards to whether *ED* is subject to amendments. Naturally, a higher expected gain, whether absolute or relative as presented in table 2, is associated with more frequent amendments in *ED*. It is also important to note that, with respect to the counterfactual case, the experiments have not considered the possibility of external factors being introduced in the hypothetical world, as demonstrated in the example of section 4.2. It is understood that the gain associated to the counterfactual case has the potential to increase greatly under such circumstances.

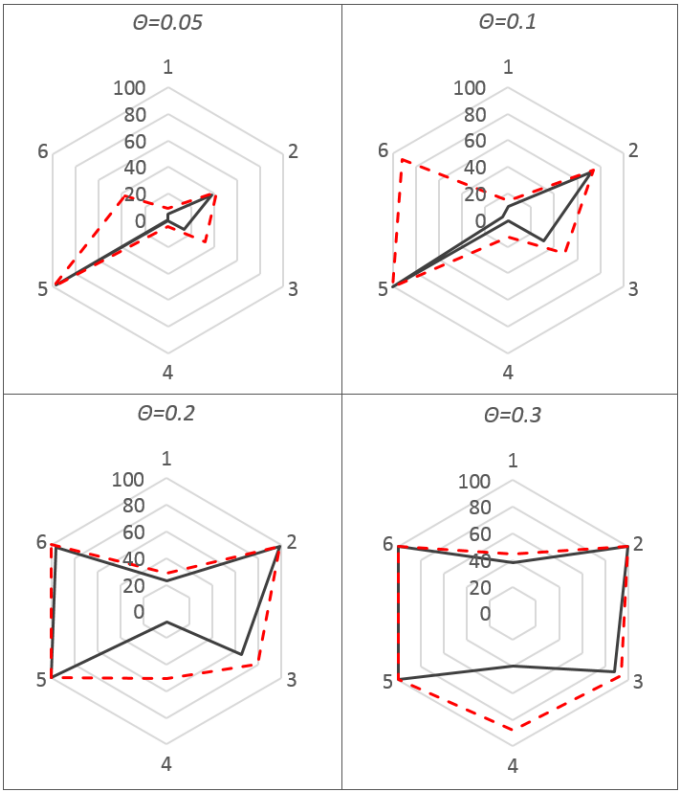


Figure 9. The average gain, when applying the method to each of the six experiments performed, in terms of *ED* for the average individual and with respect to being determined suitable for release at various threshold levels *Θ*. The solid line indicates the average model expectations prior to applying the method, whereas the dashed line indicates the shift towards an increased chance of the average individual determined as being suitable for release, after the method has been applied.

6 Concluding remarks

The concept of *VoI* analysis has been widely studied, primarily with influence diagrams, for observational and interventional cases. The novelty of our work here is to show how *VoI* can be used directly with BNs for interventional and especially counterfactual cases. Our application of *VoI* is also somewhat different from standard use. Typically, *VoI* is used for a) determining the amount a decision maker would be willing to pay for further information, and b) prioritising unknown factors for acquiring additional information based on their impact against a desired utility value or probability distribution.

In this paper, we used *VoI* analysis to determine whether a decision about an interventional action could be subject to amendments on the basis of some incomplete information within the model, and whether it would be worthwhile for the decision maker to seek further information prior to suggesting a decision. We have described a method to incorporate *VoI* analysis into BNs to enhance decision analysis in medical applications concerned with interventional actions. That is, we have also shown how the underlying principle of *VoI* analysis becomes useful, not only when decision making is subject to amendments, but also when it is not. Knowing that a decision outcome is independent of one or more unknown risk factors saves us from the trouble of seeking information about the particular set of risk factors. Further, we have also extended the assessment of this implication to the counterfactual case and demonstrated how answers about interventional actions are expected to change when some unknown factors become known, and how useful this becomes in forensic medical science.

The process of *VoI*, which can be seen as an extension to sensitivity analysis [17, 37], can be automated[8] to examine amendments in the *EDs* on the basis of some relevant interventional actions of interest. In contrast, interventional and especially counterfactual BNs will typically require careful reconstruction, from observational BNs, that might not always be consistent and in some cases might require expert contribution (see Figs. 3, 7 and 8). While Pearl has provided the basic formal underpinning of these processes [30], they appear to suffer from some limitations and as a result, some extensions of these processes (e.g. such as from *perfect* to *imperfect* interventions as discussed in the previous sections) have been proposed [32, 40, 41]. The scalability challenges in performing such extended *VoI* analysis are being addressed in [42].

Real-world decision making is hindered by severe uncertainties, and these uncertainties are typically expected to increase greatly when the decision problem incorporates interventional and counterfactual questions. While in this paper we have focused our analysis on forensic medical sciences, the modelling process still

[8] The BN models presented in this paper do not incorporate continuous variables for unobserved factors that can be manipulated. Performing *VoI* on continuous variables typically requires some complex approximations. In [38] we demonstrate how to perform *VoI* analysis in BNs using *Dynamic Discretisation* [39].

applies to any other similar real-world problem that incorporates interventional and counterfactual Bayesian simulations.

Acknowledgements

We acknowledge the financial support by the European Research Council (*ERC*) for funding this research project, ERC-2013-AdG339182-BAYES_KNOWLEDGE, and *Agena Ltd* for software support.

References

[1] Raiffa, H., Schlaifer, R. (1961), Applied Statistical Decision Theory. Harvard University Graduate School of Business Administration, Cambridge, Massachusetts.

[2] Pflug, G. C. (2006). A value-of-information approach to measuring risk in multi-period economic activity. *Journal of Banking & Finance*, 30(2), 695-715.

[3] Hahn, G. J., & Kuhn, H. (2012). Value-based performance and risk management in supply chains: A robust optimization approach. *International Journal of Production Economics*, 139(1), 135-144.

[4] Sculpher, M., & Claxton, K. (2005). Establishing the Cost-Effectiveness of New Pharmaceuticals under Conditions of Uncertainty—When Is There Sufficient Evidence? *Value in Health*, 8(4), 433-446.

[5] Baio, G. (2012). *Bayesian methods in health economics.* CRC Press, Boca Raton.

[6] Claxton, K., & Posnett, J. (1996). An economic approach to clinical trial design and research priority-setting. *Health Economics*, 5(6): 513-524.

[7] Claxton, K., & Thompson, K. M. (2001). A dynamic programming approach to efficient design of clinical trials. *Journal of Health Economics*, 20: 797-822.

[8] Halpern, J., Brown, Jr. B. W., Hornberger, J. (2001). The sample size for a clinical trial: a Bayesian-decision theoretic approach. *Statistics in Medicine*, 20: 841-858.

[9] Eckermann, S., & Willan, A. R. (2007). Expected value of information and decision making in HTA. *Health Economics*, 16: 195-209.

[10] Kikuchi, T., Pezeshk, H., Gittins, J. (2008). A Bayesian cost-benefit approach to the determination of sample size in clinical trials. *Statistics in Medicine*, 27(1): 68-82.

[11] Ramsey, S. D., Blough, D. K., & Sullivan, S. D. (2008). A forensic evaluation of the national emphysema treatment trial using the expected value of information approach. *Medical Care* 46: 542–548.

[12] Willan, A. R., Eckermann, S. (2010). Optimal clinical trial design using value of information methods with imperfect implementation. *Health Economics*, 19: 549–561.

[13] Griffin, S., Welton, N. J., & Claxton, K. (2010). Exploring the research decision space: the expected value of information for sequential research designs. *Medical Decision Making*, 30(2): 155-162,

[14] Braithwaite, S., & Scotch, M. (2013). Using value of information to guide evaluation of decision supports for differential diagnosis: is it time for a new look? *BMC Medical Informatics & Decision Making*, 13: 105.

[15] Sadatsafavi, M., Bansback, N., Zafari, Z., Najafzadeh, M. & Marra, C. (2013). Need for Speed: An Efficient Algorithm for Calculation of Single-Parameter Expected Value of Partial Perfect Information. *Value in Health*, 16, 438-448.

[16] Strong, M. & Oakley, J. E. (2013). An Efficient Method for Computing Single-Parameter Partial Expected Value of Perfect Information. *Medical Decision Making*, 33, 755-766.

[17] Strong, M., Oakley, J. E., & Brennan, A. (2014). Estimating Multiparameter Partial Expected Value of Perfect Information from a Probabilistic Sensitivity Analysis Sample: A Nonparametric Regression Approach. *Medical Decision Making*, 34(3): 311-326.

[18] Yokota, F., & Thompson, K. M. (2004). Value of Information Literature Analysis: A Review of Applications in Health Risk Management. *Medical Decision Making*, 24: 287-298.

[19] Lu, T.-C., & Druzdzel, M. J. (2002). Causal models, value of intervention, and search for opportunities. In Gamez, J. A., & Salmeron, A., eds., *Proceeding of the First European Workshop on Probabilistic Graphical Models* (PGM'02), 108–116.

[20] Shea, K., Tildesley, M. J., Runge, M. C., Fonnesbeck, C. J., & Ferrari, M. J. (2014). Adaptive Management and the Value of Information: Learning Via Intervention in Epidemiology. *PLoS Biology*, 12:(10).

[21] Coid, J. W., Ullrich, S., Kallis, C., Freestone, M., Gonzalez, R., Bui, L. et al. (2014). Improving Risk Management in Mental Health Services. *British National Institute for Health Research (NIHR)*, July 2014.

[22] Constantinou, A. C., Freestone, M., Marsh, W., & Coid, J. (2014). Causal inference for violence risk management and decision support in forensic psychiatry. *Decision Support Systems*, 80: 42-55. Draft available at: http://constantinou.info/downloads/papers/DSVM-MSS.pdf (Accessed July 18, 2015)

[23] Constantinou, A. C., Freestone, M., Marsh, W., Fenton, N. & Coid, J. (2015). Risk assessment and risk management of violent reoffending among prisoners, *Expert Systems with Applications*, 42(21): 75111-7529.

[24] Fenton, N.E., & Neil, M. (2012). *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press, Boca Raton.

[25] Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. New England Journal of Medicine, 299, (18) 999-1001.

[26] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann Publishers.

[27] Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques – Adaptive Computation and Machine Learning*. The MIT Press, Cambridge, Massachusetts.

[28] Hubbard, D. (2007*). How to Measure Anything: Finding the Value of Intangibles in Business*. John Wiley & Sons, New Jersey.

[29] Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007) Causal reasoning through intervention. In *Causal learning: Psychology, philosophy and computation*, ed. A. Gopnik & L. Schulz. Oxford University Press, Oxford.

[30] Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. 2nd edition, Cambridge University Press, Cambridge, UK.

[31] Korb, K., Hope, L., Nicholson, A., & Axnick, K. (2004). Varieties of causal intervention. In *Pacific Rim Conference on AI, 2014.*

[32] Eaton, D., & Murphy, K. (2007). Exact Bayesian structure learning from uncertain interventions. In *AI & Statistics*, 2007.

[33] Constantinou, A. C., Marsh, W., & Fenton, N. (2015). From complex questionnaire and interviewing data to intelligent Bayesian models. Under review.

[34] Dawid, A. P. (2000). Causal Inference without Counterfactuals. *Journal of the American Statistical Association*, 95: 407-424.

[35] Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, Cambridge.

[36] Dawid, P. (2012). The Decision-Theoretic Approach to Causal Inference. In *Causality: Statistical Perspectives and Applications*. John Wiley & Sons, Chichester, UK; chapter 4, pp. 25-42.

[37] Felli, J. C., & Hazen, G. B. (1998). Sensitivity analysis and the expected value of perfect information. *Medical Decision Making*, 18(1): 95-109.

[38] Yet, B., Constantinou, A., Fenton, N., & Neil, M. (2015). Partial Expected Value of Perfect Information with Dynamic Discretisation. Under review.

[39] Neil, M., Marquez, D., & Fenton, N. (2010). Improved Reliability Modeling using Bayesian Networks and Dynamic Discretization. *Reliability Engineering & System Safety*, 95(4), 412-425.

[40] Lucas, A., & Kemp C. (2012). A unified theory of counterfactual reasoning. *In Proc. 34th Annu. Meet. Cogn. Sci. Soc.*, ed. Miyake, N., Peebles, D, & Cooper, R.P, pp. 707–12. Austin, TX: Cogn. Sci. Soc.

[41] Sloman, S. A., & Lagnado, D. (2015). Causality in Thought. *The Annual Review of Psychology*, 66: 223-247.

[42] Fenton, N. (2014). Effective Bayesian Modelling with Knowledge Before Data (BAYES-KNOWLEDGE). European Research Council (ERC) Advanced Grant. http://www.eecs.qmul.ac.uk/~norman/projects/B_Knowledge.html (Accessed July 18, 2015).

Appendix A: CPTs for the *interventional* BN model of Figure 7

Table A.1. CPT for Delusions (*D*), Anger (*A*), Substance misuse (*S*), and Background of extreme violent behaviour (*B*).

Delusions		**Anger**		**Substance misuse**		**Back. of extreme violent beh.**	
False	0.6	False	0.4	False	0.2	False	0.8
True	0.4	True	0.6	True	0.8	True	0.2

Table A.2. CPT for Delusions post treatment (*Dt*).

Delusions	False		True	
Treat. for mental illness	False	True	False	True
False	1	1	0	0.6
True	0	0	1	0.4

Table A.3. CPT for Anger post treatment (A*t*).

Anger	False		True	
Treatment for anger	False	True	False	True
False	1	1	0	0.3
True	0	0	1	0.7

Table A.4. CPT for Substance misuse post treatment (S*t*).

Substance misuse	False		True	
Substance misuse treat.	False	True	False	True
False	1	1	0	0.7
True	0	0	1	0.3

Table A.6. CPT for Disinhibition (*Di*).

Substance misuse post treat.	False	True
False	1	0.2
True	0	0.8

Table A.7. CPT for Violent ideation (*Vi*).

Back. of extreme violent beh.	False		True	
Delusions post treatment	False	True	False	True
False	1	0.8	0.6	0.1
True	0	0.2	0.4	0.9

Table A.8. CPT for Uncontrolled aggression (*U*).

Anger post treatment	False		True	
Disinhibition	False	True	False	True
False	0.9	0.6	0.3	0.1
True	0.1	0.4	0.7	0.9

Table A.9. CPT for Violence (*V*).

Violent ideation	False		True	
Uncontrolled aggression	False	True	False	True
p(Violence=*true*)	0.05	0.3	0.55	0.75

Appendix B: New/Revised CPTs for the *actual-world* BN model of Figure 8.

Note that CPTs for nodes *Cultural, ethnic, or religious incentives* (*C* & *C'*) and *Spiritual care* (*Sc*) are not provided since these variables are observable (i.e. set to *true*), in the example, without any parent nodes. The outcome of interest is retained whatever the CPT values provided.

Table B.1. CPT for Violent ideation (*Vi* & *Vi'*), where *CERI* is *Cultural, ethnic, or religious incentives*, *BEVB* is *Background of extreme violent behaviour*, and *DPT* is *Delusions post treatment*.

CERI	False				True			
BEVB	False		True		False		True	
DPT	False	True	False	True	False	True	False	True
False	1	0.8	0.6	0.1	0.5	0.3	0.2	0.01
True	0	0.2	0.4	0.9	0.5	0.7	0.8	0.99

Table B.2. CPT for Violence (*V* & *V'*).

Violent ideation	False		True	
Uncontrolled aggression	False	True	False	True
p(Violence=*true*)	0.05	0.3	0.75	0.9

Table B.3. CPT for Cultural, ethnic, or religious incentives post care (*Cc*).

Cultural, ethnic, or religious incentives	False		True	
Spiritual care	False	True	False	True
False	1	0.8	0.6	0.1
True	0	0.2	0.4	0.9