EXPLAINING INFERENCE ON A POPULATION OF INDEPENDENT AGENTS USING BAYESIAN NETWORKS

by

Peter Šutovský

M.S., Comenius University, Bratislava, 1992

Submitted to the Graduate Faculty of

the School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Peter Šutovský

It was defended on

July 29, 2013

and approved by

Gregory F. Cooper, Professor, Biomedical Informatics and Intelligent Systems

Marek J. Druzdzel, Associate Professor, School of Information Sciences

Roger R. Flynn, Associate Professor, School of Information Sciences

Michael Lewis, Professor, School of Information Sciences

Milos Hauskrecht, Associate Professor, Computer Science

Dissertation Director: Gregory F. Cooper, Professor, Biomedical Informatics and Intelligent Systems

EXPLAINING INFERENCE ON A POPULATION OF INDEPENDENT AGENTS USING BAYESIAN NETWORKS

Peter Šutovský, PhD

University of Pittsburgh, 2013

The main goal of this research is to design, implement, and evaluate a novel explanation method, the hierarchical explanation method (HEM), for explaining Bayesian network (BN) inference when the network is modeling a population of conditionally independent agents, each of which is modeled as a subnetwork. For example, consider disease-outbreak detection in which the agents are patients who are modeled as independent, conditioned on the factors that cause disease spread. Given evidence about these patients, such as their symptoms, suppose that the BN system infers that a respiratory anthrax outbreak is highly likely. A public-health official who received such a report would generally want to know *why* anthrax is being given a high posterior probability. The HEM explains such inferences. The explanation approach is applicable in general to inference on BNs that model conditionally independent agents; it complements previous approaches for explaining inference on BNs that model a single agent (e.g., for explaining the diagnostic inference for a single patient using a BN that models just that patient). The hypotheses that were tested are: (1) the proposed explanation method provides information that helps a user to understand how and why the inference results have been obtained, (2) the proposed explanation method helps to improve the quality of the inferences that users draw from evidence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to take this opportunity to thank people who influenced my dissertation research.

First, and foremost, I would like to thank my advisor, Gregory Cooper. He has guided my research for many years. I wish to express my deepest gratitude to him for his patient supervision, his enthusiastic support, and valuable insights. His advice was not only confined to my research, has been always invaluable, always well thought out, always clear, and concise, and always supported by a strong argument. He has set an example for me that I will always strive to follow. I have learned not only from his profound professional knowledge but also from his giving personality and working attitude. He is one of the nicest and most helpful advisors that a graduate student could have.

Also, I would like to thank Marek Druzdzel who was advising me in my research during my first years in the School of Information Sciences while I was working in the Decisions Systems Laboratory (DSL) at the University of Pittsburgh and then during my work on my dissertation. He guided me in the writing of my first research paper. Later during my work on my dissertation he gave me very useful suggestions. I often asked him for help or advice and he was always ready to help. His professional and personal advice was always most valuable to me.

In addition to Greg and Marek I would like to thank the other members of my dissertation committee - Roger Flynn, Milos Hauskrecht, and Michael Lewis – for all their support, and most of all, for their patience. I appreciate their insightful and helpful comments on my dissertation work.

I would like to thank my colleagues at Cranfield University, Adam Zagorecki, Ken McNaught, and Piers MacLean, and my former colleague from DSL, Mark Voortman, for their useful suggestions and advice and helping me to organize this research study.

I am also thankful to Garrick Wallstrom and Trevor Ringrose, with whom I consulted.

I would like to thank Hassan Karimi who supported and guided me in my research during my first year at the School of Information Sciences and who also involved me in several of his research projects during my later years.

I am grateful to my colleagues in DSL at the School of Information Sciences and in the Bayesian Biosurveillance group in the Department of Biomedical Informatics for creating such friendly and stimulating places to work.

I am very grateful for the financial support from School of Information Sciences at the University of Pittsburgh and support by a grant from the National Science Foundation (NSF IIS-0325581).

Finally, I would like to thank my parents, my brother, and my grandparents for their love and patience through my time at the University of Pittsburgh. Though we spend very limited time together during the past years, their support been always been a source of energy and optimism throughout my doctoral studies.

1.0 INTRODUCTION

The importance of an explanation facility in intelligent systems has been recognized for some time. Several studies have experimentally confirmed the positive impact of explanation on learning (Berry and Broadbent, 1987; Moffitt, 1989; Druzdzel and Henrion, 1990), on belief in a system's conclusions (Everett, 1994; Ye and Johnson, 1995), and on the accuracy of decision making (Berry and Broadbent, 1987; Suermondt and Cooper, 1993). However, to my knowledge only one experimental study has evaluated the impact of explanation of inference in a Bayesian network on decision making and on quality of prediction. (Suermondt and Cooper, 1993). This study showed that decision making can be improved by appropriate explanation (Suermondt and Cooper, 1993). Previous research on explanation in Bayesian networks has provided methods enabling generation of an explanation in arbitrary Bayesian networks (Suermondt, 1992; Haddawy et al., 1994; Chajewska and Draper, 1998). However, as explanations in Bayesian networks are computationally expensive, they may not be feasible for large models. For example, *Bayesian networks with population of agents* (BNPA) are inherently large networks, as they represent each agent in the population with its own Bayesian subnetwork. These networks are useful in situations in which we want to learn something about a population based on information about the agents in the population. In disease-outbreak detection (also known as biosurveillance), for example, the agents are often people reporting their symptoms when admitted to the hospital. Hence agent-based Bayesian networks may contain large numbers of variables and findings. Nevertheless, while using existing explanation methods for BNPAs might not be practical (or even possible), the structure of agent-based Bayesian models provides an opportunity to design specific explanation methods that are computationally more efficient and provide appropriate explanation for this type of model structure.

Most of the existing methods for explanation of inference in Bayesian networks are based on assigning values to the evidence, nodes and arcs in a *Bayesian network* (BN) that represent the importance of each of them in obtaining a particular inference result from the Bayesian network. Based on the importance ranking and required level of detail, explanation methods filter out relatively unimportant components and present only important findings, nodes, and arcs as part of the explanation. Since there are interactions between different variables and findings, in principle all possible combinations should be checked. Since this is not feasible, currently existing explanation methods use heuristics to avoid having to perform a complete search. The explanation method proposed here is most closely related to the work of Druzdzel and Henrion (1990), Suermondt (1992), and Chajewska and Draper (1998). BNPA divides a large network into subnetworks and creates explanations based on subnetworks first. In order to produce explanations within the subnetworks, existing methods can be used to complete explanations. Although there are various explanation methods for inference in a Bayesian network, only a few have been evaluated by users (Suermondt and Cooper, 1993; Druzdzel and Henrion, 1990). While Suermondt and Cooper (1993) studied effect of explanation on subjects' decision making, predictions and confidence, Druzdzel and Henrion (1990) studied effect of explanation on subjects' learning, on improvement of subjects' insight into system's reasoning.

I will evaluate a new explanation method to verify its usefulness for users and to obtain feedback which may lead to improvement of the method.

1.1 SPECIFIC AIMS

In this dissertation research, I develop, implement and evaluate a new method to explain inference in Bayesian networks representing a population of independent agents. I evaluate the influence of these explanation methods on the quality of human inference from evidence.

BNPAs model each agent in the population using the agent's subnetwork (Figure 1). These subnetworks are connected to the common subnetwork, which consists of nodes representing the cumulative characteristics of the whole population. There can be direct

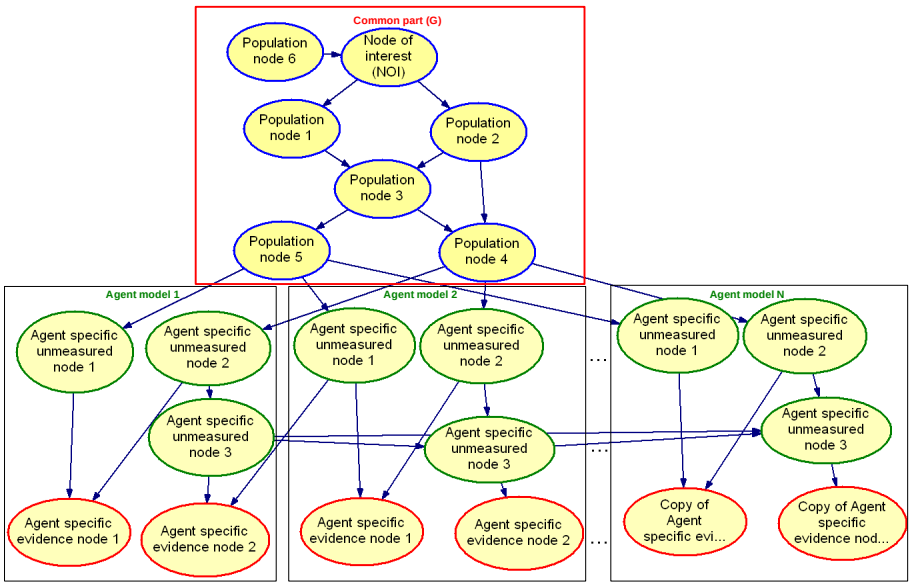


Figure 1: Agent-based Bayesian networks with interaction between agents in population

interaction between agents in the population in an agent-based BN, and these interactions are represented by a directed arc between some variables of different subnetworks of agents (see Figure 1). This study is focused on BNPAs without direct interaction between agents in a population (Figure 2). Sections 4.1.2 and 4.2.1 provide several examples in which such assumptions are reasonable. PANDA-CDCA (Cooper et al., 2006) is one instance of such a network (Section 4.1.2). I refer to this type of network as a *Bayesian Network with a Population of Independent Agents* (BNPIA), since the agents in the population will be independent of one another if we condition on all the factors that make them dependent, such as population nodes 4 and 5 in Figure 2.

Specifically, my aims in the dissertation research were as follows :

1. To develop and implement explanation methods capable of providing explanation of inference for BNPIA.
2. To evaluate the influence of the explanation provided by the method on the inference of users of a simple biosurveillance system in a laboratory setting. Synthetic data will be used for evaluation.

Figure 2: Agent-based Bayesian networks without interaction between agents in population

I hypothesize that the explanations provided to users of an agent-based BN by the proposed explanation methods will elucidate the system's inferences and thereby improve the user's inferences from evidence about the population of agents and improve the user's confidence about his or her inferences.

2.0 DEFINITIONS AND BACKGROUND

People have to make many personal and professional decisions every day. Nearly every activity requires decision making at some point. Due to incomplete information, many of these decisions must be made under a condition of uncertainty. It has been shown that human decisions made under uncertainty are often suboptimal, biased and inconsistent with probability and decision theories (Tversky and Kahneman, 1974). Friedman et al. (1999) experimentally confirmed that a decision support system can improve diagnostic accuracy. One way to support consistency in decision making is to use decision aids based on probability, statistics, and decision theory. Various methodologies have been developed in the field of artificial intelligence (AI) to deal with uncertain information, including fuzzy logic (Zadeh, 1986, 1996, 1989), certainty factors (David McAllister in the mid-1980s), the Demster-Shafer theory (Dempster, 1968; Shafer, 1976), and graphical probabilistic models (Pearl, 1988; Neapolitan, 1990).

Until the mid-1980s, the probabilistic approach to building decision support systems was considered to be impractical by the mainstream AI community due to the large number of parameters required and the computational complexity of inference. Later theoretical developments, mainly the employment of conditional independence (Charniak, 1983; Duda et al., 1990; Kim and Pearl, 1983) and local computation (Pearl, 1982), helped to address these problems. However, employment of conditional independence for complex models requires efficient representation of independence. Graphical models are well-equipped to deal with this problem since they efficiently represent qualitative relationships among variables (see Chapter 2.3 for more details).

Today, numerous AI systems use probability, statistics and decision theory to cope with decisions in an uncertain world. Tools based on probabilistic methods have a substantial

advantage over those based on other types because they provide an assessment of uncertainty that meshes with decision theory (Pearl, 1988; Neapolitan, 1990).

Graphical models are currently one of the most successful tools for modeling uncertainty and are widely applied in pattern recognition (Frey, 1998), tutoring systems (Schulze et al., 2000; Conati et al., 1997), user interfaces (Horvitz et al., 1998; Horvitz and Barry, 1995), information retrieval (Fung and del Favero, 1995), machine learning, aircraft diagnostics (Kipersztok and Wang, 2001), locomotive diagnostics (Przytula and Thompson, 2000), diagnosis and control of autonomous vehicles (Madsen et al., 2004), financial operational risk assessment (Neil et al., 2005), industrial planning (Gebhardt et al., 2006), ecology (Zhu and Deshmukh, 2003), genetics (Segal et al., 2003; Bulashevska et al., 2004), biosurveillance (Cooper et al., 2006), and medical diagnosis (Shwe et al., 1991; Lacave and Díez, 2003). Graphical models are graphical representations of probabilistic structures and functions representing local probabilities that are used to derive joint probability distribution. A graphical model is the result of a marriage between graph theory and probability theory. Graph theory provides graphical models with efficient representation and algorithms, while probability theory provides a solid theoretical foundation for modeling under uncertainty. The most popular types of graphical models are the *Markov random field* (MRF) and the *Bayesian belief network* (BBN). The BBN has been one of most successful modeling tools applied in AI to practical problems within the last 15 years.

2.1 EXPLANATION IN INTELLIGENT SYSTEMS

As intelligent systems became more frequently applied to real world problems, more and more people who were not domain experts began to use them. These computer programs were created to imitate human experts, and as an important feature of human experts is their ability to communicate their knowledge and reasoning, users expected this same ability from expert systems. However, experiments performed as part of the MYCIN project (Shortliffe, 1976) showed that physicians were very reluctant to accept recommendations from a computer if they did not understand the reasoning that led to the result (Buchanan and

Shortliffe, 1984). Tests of knowledge-based systems have also shown that detailed explanations are very important for the system's success (Everett, 1994; Moffitt, 1989; Gault, 1994; Mao, 1995; Berry and Broadbent, 1987). Users do not blindly trust the results provided by expert systems; they need to understand the reality that the model represents. Therefore, whether a user will or will not accept the system depends not only on the quality of the conclusion itself but also on an appropriate explanation for the conclusion (Swartout and Moore, 1993). In other words, understanding how and why a system reached a particular conclusion helps the user to evaluate the model. Explanation also helps the user to determine whether the conclusions of the system are reasonable given the evidence. Moreover, in addition to helping the end user to understand the conclusion of the intelligent system, explanation may also be a useful tool for debugging the system, since it can aid in verifying the domain model's validity and in detecting possible inconsistencies (Lacave et al., 2001).

2.2 DEFINITION OF EXPLANATION

According to The Philosopher's Dictionary (Martin, 1994), "An explanation answers the question 'why' and provides understanding; perhaps it also provides us with the abilities to control, and to predict (and retrodict) the world. ... One (but only one) sort of explanation is Causal: we explain something by saying what its causes are. Sometimes, instead, we explain by telling what something is made of, or by giving reasons for human Actions (but see Reasons / Causes), as in some explanations in history." The object of the explanation can be "concepts, causes, or effects, procedures/rules (e.g. evacuation procedures), purposes, objectives, relationships, and processes" (Wragg and Brown, 1993). Explanation is usually defined in the context of description, comprehension, prediction and causality. In general, the purpose of explanation is to clarify something and make it understandable.

2.3 BAYESIAN NETWORKS

One of the strongest arguments against using probability theory in decision-support systems is that it appears to be unfeasible with respect to representing knowledge probabilistically without simplifying assumptions about dependencies among the variables. Graphical models, however, solved this problem and allow us to represent arbitrary dependencies among variables.

A *Bayesian network* (BN) is an important graphical modeling tool for domains that involve uncertainty. A BN is also known as a *Bayesian belief network* (BBN), a *causal probabilistic network*, a *directed Markov field,* and, with an additional structure, an *influence diagram*. Detailed information about BNs can be found in Howard and Matheson (1981), Pearl (1988) and Neapolitan (1990). Before I proceed to the description of some key concepts of Bayesian networks I will define several terms from graph theory which I refer to in the text.

Definition 1. *A path in a graph is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence.*

Definition 2. *A* cycle *is a path such that the start vertex and the end vertex are the same.*

Definition 3. *A* directed graph *or* digraph *G is an ordered pair $G := (V, A)$, where* V *is a set, whose elements are called vertices or nodes and* A *is a set of ordered pairs of vertices, called directed edges, arcs, or arrows.*

Definition 4. *A* tree *is a graph in which any two vertices are connected by exactly one path.*

An important concept for understanding a BN is *joint probability distribution* (JPD) given by Definition 5.

Definition 5. *Joint probability distribution* (JPD) *of variables $X_1, X_2, \ldots, X_N$, is probability distribution with probabilities defined for every vector $r = (r_1, r_2, \ldots, r_N)$ in Cartesian state space $\Omega = R_1 \times R_2 \times R_3 \times \ldots R_N$ of variables $X_1, X_2, \ldots, X_N$.*

A Bayesian network is represented by two main components: a graph (Figure 3) and local probability distributions (Table 1). A BN efficiently deals with two main problems

in applied mathematics and engineering: complexity and uncertainty. Being a graphical model for describing probabilistic relationships among domain variables, a BN is capable of efficiently representing a JPD over the random variables representing nodes of a graph.

The graph is the qualitative component of BN, locally representing the relationships among domain variables. The graph of a BN is a directed acyclic graph (DAG), which means that it cannot contain cycles, that is, closed loops of directed links. The graph consists of nodes/variables and directed arcs which connect nodes. The arcs are directed from parents to children. The arc expresses a dependency of a child node on a parent node. Figure 3 presents part of the coma network (Cooper, 1984) as an illustration. For example the node 'Metastatic Cancer' is parent of the nodes 'Serum calcium' and 'Brain Tumor'. The DAG represents independences among the variables in a human-friendly way, thereby providing an intuitive interface which makes model building and debugging easier for domain experts.

Every node is associated with a *conditional probability distribution* (CPD). In general, a BN network can contain continuous variables, discrete variables or both. For simplicity, I will be discussing BNs with discrete variables (also known as discrete Bayesian networks). In the case of discrete BNs, the CPD consists of a set of distributions. One distribution is defined for each configuration of parents of the node. If the node does not have any parent, a prior probability distribution is defined for the node. These probabilities are called local probability distributions. Later I will show how joint probability distribution can be calculated using information about independences in the graph and local probability distributions.

A central idea of the BN is modularity. The BN is a complex system which is built of simpler components due to a concept of conditional independence. Conditional independence on the local level is represented by a missing arc between two nodes. Graphical representation encodes independences among variables efficiently and thus allows representation of the JPD with fewer parameters. A probabilistic interaction among the variables in the network is specified by the interactions of each node with its neighbors (see Table 1). Statistical conditional independence is related to d-separation in the graph.

Definition 6. *Two nodes are* d-separated *by the set of nodes Z if (1) in every directed path from X to Y there is at least one node from Z variables, or (2) every node which has a*

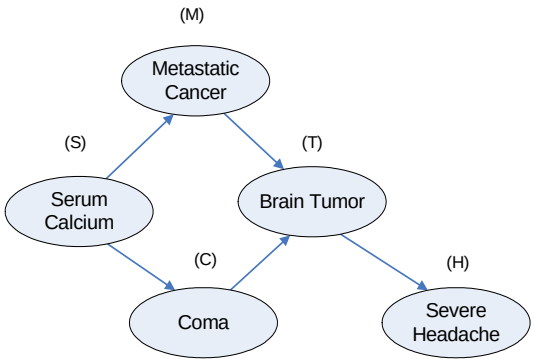


Figure 3: Bayesian network example.

directed path to both X and Y has at least one node from Z on the path to X and on the path to Y, or (3) no node which has directed path from X and from Y is in Z.

Example 1. *Some d-separations that can be found for the BN in Figure 3: nodes* M and C are *d-separated given the nodes* S *and* T*; nodes* M *and* H *are d-separated given node* T*; nodes* S *and* T *are d-separated given the node* M*, nodes* C *and* H *are d-separated given the node* T.

If two nodes are d-separated, given the set of evidence nodes Z, nodes X,Y are conditionally independent given the subset of evidence nodes Z. The conditional independence represented by a graph structure is utilized to represent a JPD with fewer parameters. Local probabilities quantify an interaction among the variables. The conditional probability distributions in Table 1 quantify the relationships for the variables in the network in Figure 3. For example, the value for the probability of S to be in state s1 when M is in state m1 is 0.2. As can be seen from parameters in Table 1, rather than encoding the JPD explicitly, the BN represents the JPD using local prior and conditional probability distributions, allowing us to represent the CPD with fewer parameters, shown in the Example 2.

Example 2. *All variables in our example network are binary. Therefore, without utilizing independences among variables, we would have to use* $2^5 - 1 = 32 - 1 = 31$ *parameters to represent JPD. However given the conditional independences among variables that are*

Table 1: Local distributions for Bayesian network in Figure 3.

P(M)	
M	
m1	0.2
m2	0.8

P(S|M)		
$S \diagdown M$	m1	m2
s1	0.2	0.8
s2	0.8	0.2

P(T|M)		
$T \diagdown M$	m1	m2
t1	0.2	0.05
t2	0.8	0.95

P(H|T)		
$H \diagdown T$	t1	t2
h1	0.8	0.6
h2	0.2	0.4

P(C|S,B)				
S	s1		s2	
$C \diagdown B$	b1	b2	b1	b2
c1	875	0.75	0.75	0.5
c2	0.125	0.25	0.25	0.5

represented by a graph, we can represent the same JPD using only $1 + 2 + 2 + 4 + 2 = 11$ *parameters. Ability to specify JPD with the fewer parameters makes model building easier.*

With respect to the notation used in this paper, a capital letter denotes a random variable, and a lower case letter denotes the state of the variable. A bold capital letter represents a set of variables and a bold lower case letter denotes the values of a set of variables. I use $\boldsymbol{Pa}(X_i)$ to denote a set of parent nodes for a node. If a node has no parents, $\boldsymbol{Pa}(X_i)$ is the empty set. I use the terms "variable" and "node" interchangeably to refer to the node in the graph.

One important concept when dealing with Bayesian networks is the Markov condition.

Definition 7. *The* Markov condition *for a Bayesian network states that any node in a Bayesian network is conditionally independent of its non-descendants, given its parents.*

Since a BN is a DAG, the nodes of a BN can be well-ordered. Well-ordering of the nodes in a BN means that parents of the node X_j have an index lower than j. Using that we can apply a chain rule of probability and we can write a JPD as:

$$P\left(X_i, \ldots, X_n\right) = \prod_{i=1}^{n} P\left(X_i \mid X_{i-1}, \ldots, X_n\right).$$

Because of the Markov condition (see Definition 7) we can write the chain rule for a Bayesian network as:

$$P\left(X_i, \ldots, X_n\right) = \prod_{i=1}^{n} P\left(X_i | \mathbf{Pa}\left(X_i\right)\right).$$

Thus, in the case of the BN shown in Figure 3 we can write:

$$P\left(M, S, T, H, C\right) = P\left(M\right) P\left(S|M\right) P\left(T|M\right) P\left(H|T\right) P\left(C|S, T\right).$$

In general, if we have n nodes, and each node is binary and has at most k parents, the number of parameters that needs to be known is at most $n \cdot 2^k$ instead of $2^n - 1$. If a graph is sparse (i.e. $k \ll n$), we can gain significant savings in the number of parameters which we need to obtain and store. A BN may contain discrete, continuous or both kinds of variables. There is ongoing research in the field of continuous BNs (BNs with continuous variables) and hybrid BNs (BNs with discrete and continuous variables) (Lauritzen, 1992; Koller et al., 1999; Lerner et al., 2001; Nachman et al., 2004). However, the BNs with discrete variables are most frequently used in practical applications. Therefore, from now on, I will be dealing only with discrete BNs.

3.0 OVERVIEW OF EXPLANATION METHODS

3.1 EXPLANATION IN BAYESIAN NETWORKS

Lacave and Diez (2002) reviewed research on explanation methods for Bayesian networks and organized them into categories. I will use their categorization of explanation methods in this paper, which is presented in Figure 4. As the figure shows, they based their organization on three main properties: content, communication, and adaptation. This paper is concerned mainly with the content of explanation. Lacave and Diez (2002) divided the category of content further, based on the focus of explanation, purpose of explanation, level of explanation and causality. Literature on explanation methods for Bayesian networks recommends that three basic aspects should be explained: (1) the evidence that was propagated, (2) the knowledge base model, and (3) the reasoning process by which results were obtained from the evidence using knowledge base. The reminder of this chapter provides a review of representative, although not exhaustive research in these areas.

3.2 EXPLANATION OF REASONING

3.2.1 Introduction

Explanation of reasoning (a.k.a. dynamic explanation) endeavors to describe comprehensibly the process by which a conclusion is obtained. In a Bayesian network, this means explaining which individual findings have the most influence on inference results and what the most important inference paths between findings E and a node of interest D are (Chajewska and

Draper, 1998; Haddawy et al., 1994; Suermondt, 1992; Madigan et al., 1997). I am using word instantiation to represent an assignment of value to the variable. I will refer to the values $e_1, \ldots, e_N$ to which the variables $E_1, \ldots, E_N$ have been instantiated as to the configuration of variables $E_1, \ldots, E_N$. Chajewska and Draper (1998) formally define predictive explanation based on Suermondt's (1992) concept of explanation of inference. I have modified the original definition of Chajewska and Draper (1998) to make it more comprehensible:

Definition 8. A *predictive explanation* X of a change in probability distribution over the node of interest D caused by the evidence δ_B (instantiation of evidence variables Δ after the change) received for the set of nodes Δ, $D \notin \Delta$, with respect to the prior distribution of the variable D for the prior configuration δ_A of the evidence nodes in Δ, is a conjunction $\delta_X \wedge P_X$ where:

- δ_X is an instantiation of nodes in Δ such that the subset of nodes $\Delta_X \subset \Delta$ is set to the values they assume in δ_B and the complementary subset of nodes $\Delta - \Delta_X$ is set to the values they assume in δ_A.
- P_X is a subset of the undirected path P linking D to the nodes in Δ_X.

Set Δ_X is called the explanation set.

Example 3. *Assume a Bayesian network with nodes A, B, C, D, E, F, G. Let D be node of interest and set of evidence nodes be $\Delta = \{A, B, C, E, F\}$. Let the configuration $\delta_A = \{A = a_1, B = b_3, C = c_1, E = e_3, F = f_2\}$ and $\delta_B = \{A = a_1, B = b_3, C = c_2, E = e_1\}$. Let the explanation set be $\Delta_X = \{E, F\}$; then instantiation of nodes in Δ is $\delta_X = \{A = a_1, B = b_3, C = c_1, E = e_1, F = \emptyset\}$. δ_X is given by configurations δ_B for nodes in Δ_x and by configuration δ_A for nodes in $\Delta - \Delta_X$. Explanation X of change of probability distribution over the node D caused by evidence δ_B with respect to δ_A is a conjunction of instantiation of δ_X and the subset of the undirected path P_X linking Δ_X to D. For the Bayesian network in Figure 5 P_X is given by*

$$P_X = \{EBAD, EBFD, EBD, EBCFD, FD, FBED, FCBED, FBD, FCBD\}.$$

Explanation that existing methods provide by default are not always concise. As they often remove irrelevant or less relevant information for explanation, called *simplification*

of explanation (Suermondt, 1992; Haddawy, Jacobson, and Kahn, 1994; Chajewska and Draper, 1998; Druzdzel and Henrion, 1990; Madigan, Mosurski, and Almond, 1997), the resulting default level of details then provided by the explanation may not match the user's expectations. However, simplification of explanation, when performed in an early stage of construction of explanation, can improve the computational efficiency of an explanation algorithm.

The process of simplification of explanation of inference can be divided into several steps. First, the most relevant findings are selected. Different measures can be used to measure relevancy of a finding (also known as quality of explanation) with respect to its influence on the distribution of the variable of interest, D, also known as the target variable. Some explanation methods (Suermondt, 1992; Chajewska and Draper, 1998; Haddawy et al., 1994) use cost function H, which measures a cost of the change in probability of the node of interest due to evidence E. Suermondt (1992) gives a nice overview of properties of various cost functions. Chajewska and Draper (1998) propose general properties that a good cost function should satisfy and suggest using two cost functions. The number of most relevant findings can be regulated by choosing a threshold for the value of cost function. This allows regulation of how much detail an explanation contains. Madigan et al. (1997), on the other hand, use weight of evidence (see Section 3.3.5) to measure influence of evidence on the variable of interest and to measure how evidence flow is restricted at each intermediate node.

After the most relevant findings have been determined, the most relevant chains of reasoning are selected. Suermondt (1992) and Druzdzel and Suermondt (1994) collected a set of methods for selecting relevant nodes for explanation for given evidence set E and the node of interest D which allow only relevant parts of the BN to be included in the explanation. First, graphical information and d-separation are used to eliminate irrelevant information. Suermondt (1992) also excludes nodes that are not on the direct path between the finding and the node of interest. In the next step information about probability distribution is used. A cost function is used to measure how relevant nodes and arcs are to changes in the probability distribution of the node of interest due to propagation of relevant evidence. A comprehensive description of the process can be found in Suermondt's thesis (Suermondt,

1992). Later, other similar explanation methods were designed. Madigan et al. (1997) proposed a graphical explanation method that works if there is only one path between each piece of evidence and node of interest. This condition is satisfied by a BN with a tree structure (Definition 4). Since there is only one path between a finding and a node of interest, a selection of relevant chains is not needed. In 1990, Madigan and Mosurski designed an algorithm which converted Berge networks to tree. This allows extension of the application of the graphical explanation method to Berge networks. The conversion of a Berge network into a tree can be viewed as a simplification of explanation. Explanation simplification, described by Suermondt (1992), Druzdzel and Suermondt (1994), Haddawy et al. (1994) and Chajewska and Draper (1998), can be performed before any other method of explanation is applied.

I organize methods for explanation of inference according to their approach to constructing explanations. The categories include: methods for qualitative explanation (see Section 3.2.3), methods for explanation of belief propagation based on the belief propagation algorithm developed by Pearl (1988) (see Section 3.2.4), and methods based on a quality of explanation measure (see Chapter 3.3). A scenario based explanation proposed by Druzdzel and Henrion (1990) is a separate category (see Section 3.2.2).

3.2.2 Scenario-based explanation

The scenario-based explanation designed by Druzdzel and Henrion (1990) was motivated by empirical studies which showed that people do not estimate and process uncertainty according to axioms of probability theory (Tversky and Kahneman, 1974). Several studies have shown that people tend to explain and interpret events and processes by weighting the most probable stories with scenarios that include the hypothesis in the focus (Tversky and Kahneman, 1974; Pennington and Hastie, 1988). *Scenario*, as defined by Druzdzel and Henrion (1990), is assignment of values to the variables that are relevant to a certain hypothesis and observed evidence in such way that they create coherent story. Explanation is provided as the listing of the most probable scenarios that are consistent with hypothesis and evidence. A posterior probability of the scenario is listed together with the scenario.

Scenario-based explanation does not require user knowledge of probability and is applicable mainly to causal networks. Probabilities are given numerically, and neither adaptation nor user-system interaction is described in Druzdzel and Henrion (1990).

A preliminary experimental study showed an improvement in understanding of the system's reasoning for subjects that had been practicing using a decision support system with scenario-based explanation when compared to subjects practicing using the same decision support system but without the explanation facility. However, there was no significant difference in improvement in understanding of the system's reasoning for the subjects practicing using a decision support system with the scenario-based explanation and for the subjects practicing using a decision support system with a belief propagation based explanation.

3.2.3 Qualitative explanation

Qualitative explanation is an alternative approach to explanation of reasoning, where information about change in probabilities is expressed in a qualitative way (sign of a change in belief). A commonly used qualitative representation of Bayesian networks is *the qualitative probabilistic network* (QPN) (Wellman, 1990a,b). Qualitative explanations are based on the transformation of a Bayesian network into a qualitative probabilistic network (QPN) (Druzdzel, 1996; Henrion and Druzdzel, 1991, 1990; Druzdzel and Henrion, 1993). In a QPN, numerical relationships among variables are replaced with qualitative influences and synergies. The main advantage of a QPN is the efficient propagation of a belief. Druzdzel and Henrion (1993) designed a polynomial time algorithm for qualitative belief propagation in a QPN. Although efficient inference in QPN is important advantage problem of QPN is the low precision of the results due to the limitation of describing the relationship between variables using only two indicators of influence: positive and negative. In order to avoid ambiguous results, Renooij and van der Gaag (1999) proposed using an enhanced QPN. Enhanced QPNs is able to distinguish strong and weak influences, and therefore is able to resolve some conflicts that would lead to ambiguous results (Renooij and van der Gaag, 1999; Renooij, van der Gaag, Parsons, and Green, 2000).

3.2.4 Explanation based on Pearl's belief propagation algorithm

Explanation of reasoning in polytrees The belief propagation algorithm, developed by Pearl (1988), is based on updating the belief at each node X using the messages sent by children (from the effects) $\lambda(x)$ (a.k.a diagnostic term) and parents (from the causes) $\pi(x)$ (a.k.a causal term) of the node X.

Definition 9. A *polytree network* (also known as *a singly connected network*) is a type of network in which there is at most one undirected path between any two nodes.

In *a polytree network,* the posterior probability of node X is obtained by $Belief(x) = \alpha\lambda(x)\pi(x)$ (Pearl, 1988). Sember and Zukerman (1989) developed an explanation of changes in the posterior distribution of node X which are explained using the changes in $\lambda(x)$ and $\pi(x)$.

Approximate explanation of reasoning Wiegerinck (2004) proposed a method for approximate explanation of reasoning which is based on the method of Sember and Zukerman (1989) described in Section 3.2.4. Wiegerinck's approach allows the explanation method of Sember and Zukerman (1989) to be applied to a network with loops. Wiegerinck's method first transforms the model with loops locally into a polytree. After that, the local polytree can be again decomposed into causal and diagnostic terms, which are then used for explanation in the same way proposed by Sember and Zukerman.

3.2.5 Graphical explanation of reasoning

Some software for Bayesian networks such as *BayesiaLab* (Bayesia SA.), *ELVIRA* (Lacave et al., 2001), *NETICA* (Norsys Inc.), *GeNIe* (Druzdzel, 1999) and *HUGIN* (Andersen et al., 1990) is able to display distributions of model variables graphically. These systems provide explanation of reasoning on the microlevel. The variation of probability is usually shown as a bar graph (see Figures 6 and 7). This approach is applicable to both causal and non-causal networks. Probabilities are expressed quantitatively and qualitatively.

BayesiaLab uses symbols and colors to show strength and direction of probabilistic relation between two directly connected nodes. The same graphical presentation is also used for static explanations.

ELVIRA also enables storage and display of evidence cases, navigation through them, and generation of new cases (Lacave et al., 2000, 2001). Moreover, ELVIRA enables the user to add findings sequentially and to observe the variation in probability which results from each additional finding. This allows the user to perform what-if analysis and to study the impact of each piece of evidence on the probability of a certain node. ELVIRA graphically displays the posterior probability distributions of the variables for different evidence cases (Figure 6). Nodes are conditionally colored according to the direction of the change in probability distribution due to new evidence with respect to the probability distribution for the previous evidence case. ELVIRA, however, does not use links to provide a dynamic explanation. User-system interaction is facilitated in ELVIRA by means of windows and menus. Simple adaptation is available through a precision threshold.

GeNIe, on the other hand, does provide a dynamic explanation, by using colored arcs, the thicknesses of which are proportional to the strength of a probabilistic relation between two directed nodes (Figure 7).

The graphical display of Madigan et al. (1997) uses arcs to display potential and actual strength of evidence available (see Figure 13 and Section 3.3.5). Madigan et al. (1997) also proposed an alternative way to display strength of evidence that is shown in Figure 8.

3.3 EXPLANATION OF REASONING USING QUALITY MEASURES

3.3.1 NESTOR

NESTOR, the decision support system built by Cooper (1984), was one of the first decision support systems based on a Bayesian belief network, and it also included an explanation facility. The explanation facility offered two functions: *compare* and *critique*. The command *"compare"* allowed a user to compare how two alternative diagnostic hypothesis are

supported by the current set of findings. The command "critique", unlike "compare", contrasted a selected hypothesis with all alternative hypotheses rather than a single hypothesis. Qualitative verbal explanations of causal chains between findings and hypothesis of interest were generated for the available commands. For the "compare" command, NESTOR displays change in ratio of posterior probabilities of two selected hypothesis due to different findings. Similarly, for the command "critique", NESTOR displays the ratio of posterior probabilities of THE selected hypothesis and the probabilities of all other alternative hypotheses. NESTOR also provides a description of the directed pathways between the disease etiology (hypothesis) and the findings (evidence). The explanation in NESTOR was designed for use with causal Bayesian networks. The explanation facility does not include a user model or adaptation to the user.

3.3.2 Explanations for naive Bayes

Spiegelhalter and Knill-Jones (1984) suggested using *weights of evidence* (*WOE*) for the analysis of evidence. *WOE* was originally proposed by Good (1977) as a measure of explicativity to quantify to what extent an event A explains why another event B should be believed. Spiegelhalter and Knill-Jones (1984) suggested using WOE for analysis of evidence when a naive Bayes or tree-like structure is assumed. They suggested presenting WOE available for the studied hypothesis H. The explanation was presented in the form of a balance sheet where evidence supporting evidence is on the one side of the balance sheet and the evidence contradicting the hypothesis is on the other side of the balance sheet (Figure 9).

3.3.3 PATHFINDER

PATHFINDER was one of the first expert systems to include some kind of explanation (Heckerman, 1990). PATHFINDER was built to help surgical pathologists with the diagnosis of lymph-node disease. Weight of evidence (WOE), as proposed by (Good, 1985), is used in the system as an aid to discriminate between two diseases or two disease groups d_1, d_2:

$$\frac{P\left(f_i \mid d_1, e\right)}{P\left(f_i \mid d_2, e\right)},$$

where f_i is some observation of the feature f and e is evidence. WOE reflects the degree to which probability for d_1 changes relative to d_2 due to additional evidence f_i. The system displays bar graphs with values proportional to WOE for each possible value of feature f to show how the outcomes for feature f influence relative probabilities for d_1 and d_2 given the evidence e. The presentation of the explanation is graphical, probabilities are expressed numerically and interaction of the user with the system is by means of windows, menus and dialog boxes.

3.3.4 INSITE

Early comprehensive work on explanation of inference in Bayesian networks was done by Suermondt (1992). I review his work in some detail here because it is highly relevant to my proposed research. His explanation methods focused on explanation of reasoning. Suermondt implemented his methods in the system called INSITE. Some explanation methods proposed by other researchers use a similar approach to explanations but with certain modification of INSITE explanations (Haddawy et al., 1994; Chajewska and Draper, 1998). INSITE first identifies *influential evidence* and then it finds *chains of reasoning* among the influential evidence and nodes of interest D. Suermondt (1992) uses *a cost function* $H\left(P\left(D\right);P\left(D|E\right)\right)$ to measure changes in the probability distribution of variables of interest D due to a set of *findings* E. The cost function $H\left(P\left(D|E_1\right);P\left(D|E_1,E_2\right)\right)$ can be interpreted as a measure of the quality of explanation since cost function quantifies how much evidence E_2 is worth in terms of change in the probability distribution; the greater the change, the higher the quality of the explanation including E_2 is. The cost function is also used to measure importance of nodes and arcs for propagating the evidence to the node of interest. Suermondt reviewed both utility–based and scoring–based cost functions. Among the cost functions that he compared were difference in expected utilities, expected linear error, mean squared error, absolute log–odds difference, difference of Brier scores, and cross–entropy. The cross-entropy (also known as relative entropy and information gain) in Equation 3.1 measures information gain about nodes of interest D due to evidence E.

$$H\left(P\left(D\right);P\left(D|E\right)\right)=\sum_i\left(p\left(d_i|E\right)log\frac{p\left(d_i|E\right)}{p\left(d_i\right)}\right). \tag{3.1}$$

Suermondt chose to use cross–entropy in the INSITE explanation facility. His explanation method, however, is not restricted to using cross–entropy but can use any cost function. The value of the measure is *significant* if its value is higher than a predetermined threshold θ.A formal definition of significance of the measure H is provided below.

Definition 10. *Cost of change H in probability is* significant *if and only if $H > \theta$.*

Selecting relevant findings First, INSITE identifies findings which have a significant effect on the variable of interest. *Variable of interest* is the variable for which we want to explain change of probability distribution due to new evidence. I use the term variable of interest and target variable interchangeably. Variable of interest is denoted by D. For example, in the Bayesian network in Figure 3, we want to explain why the posterior probability of the variable 'Metastatic Cancer' is high given the observed evidence. In this case, the variable 'Metastatic Cancer' is the variable of interest. In *decision support systems* (DSS), with a large *knowledge base* (KB), not all findings in the evidence set are relevant to the variable of interest. Explanation is easier to understand if it can be made simpler by focusing on relevant findings only. *Joint cost of omission* can be used to determine which findings are explanatory, sufficient and crucial. *Cost of omission* of the subset of findings $F \subseteq E$ is defined using the following cost function:

$$H^{-}(F) = H\left(P\left(D|E\right); P\left(D|E \setminus F\right)\right),$$

where "\" is a set-difference operator. Evidence F is *significant* with respect to a given threshold θ if $H^{-}(F) > \theta$. For omission of $\neg F = E \setminus F$, its cost is equal to the change from $P\left(D|E\right)$ to $P\left(D|F\right)$.

The set of finding F is sufficient if evidence in F achieves approximately the same result as evidence in E:

Definition 11. A set of findings $F \subseteq E$ is *sufficient* (to obtain inference results) if and only if $H^{-}(\neg F) \leq \theta$.

The set of findings F is explanatory if omitting F from the evidence gives substantially different results than evidence E:

Definition 12. The subset of findings $F \subseteq E$ is *explanatory* if and only if

$$H^{-}(F) > \theta\,.$$

Suermondt (1992) calls finding E_i *crucial* given the threshold θ if and only if

$$\forall F \subseteq E : \left(H^{-}(F) > \theta \wedge H^{-}(\neg F) \leq \theta\right) \Rightarrow E_i \in F.$$

Simply put, E_i is *crucial* if it is an element of every subset F of the available evidence E such that F is explanatory and sufficient to produce the change from $P(D)$ to $P(D|E)$. INSITE provides a list of sufficient and crucial evidence sets as well as graphical highlighting of crucial and relevant findings as an explanation of evidence.

The correct way to search for relevant findings would be to analyze every possible subset of findings. Suermondt (1992) refers to this as to as multi-way analysis. Unfortunately, the number of possible subsets is exponential in the number of findings. In order to account for computational complexity, INSITE first estimates the time needed for a multi-way analysis. If the multi-way analysis would take too long, the system performs a one-way analysis to select relevant findings. One-way analysis evaluates each finding E_i separately. INSITE also performs a sufficiency test that checks the reliability of the one-way analysis . The sufficiency test checks whether the joint cost of omission of all findings that are not in set of selected findings S are insignificant, $H^{-}(\neg S) \leq \theta$. If the selected set of findings does not pass the test, the one-way analysis is unreliable and selected findings S does not represent relevant evidence with required explanatory power.

Suermondt (1992) also included *a conflict detection* method in INSITE's explanation facility. This conflict detection method is based on two measures: the cost of omission and the direction of change. Direction of change is defined as follows:

Definition 13. *Direction of change* from probability distribution $P(D)$ to $P'(D)$ for the node D with the states $d_1, d_2, \ldots, d_n$ is the vector

$$Dir\left(P(D); P'(D)\right) = (dir_1, dir_2, \ldots, dir_n)$$

in which $dir_i = sign\left(p'(d_i) - p(d_i)\right)$. Possible values of dir_i are "+", "-", "0".

Finding E_i is not in conflict with the remainder of evidence if E_i does not dominate in E

$$H^{-}(E_i) \leq H^{-}(E),$$

and direction of change due to E_i is the same as direction of change due to E:

$$Dir(P(D \mid E); P(D \mid E \setminus E_i)) = Dir(P(D \mid E); P(D)).$$

On the other hand, if

$$Dir(P(D \mid E); P(D \mid E \setminus E_i)) \neq Dir(P(D \mid E); P(D)),$$

finding E_i conflicts with the remaining evidence. This conflict detection method takes into account the effect of evidence on the variable of interest. This approach detects conflict of evidence with respect to variable of interest unlike the methods of Chamberlain and Nordahl (1989) and Jensen, Olesen, and Andersent (1990) which analyze evidence independently of the variable of interest.

The next section describes another step in the construction of explanation — identification of chains of reasoning.

Chains of reasoning INSITE also finds chains of reasoning, through which findings affect the variable of interest. Inference in the network can be approximately viewed as flow of information from the evidence to the variable of interest. Generally there are multiple paths between two nodes in the Bayesian network. Only in the special case of a *simply connected* Bayesian network (also known as *polytree* Bayesian network) is there only one path between any pair of nodes. Suermondt (1992) developed a method that identifies sections of the network that are relevant to transmission of important findings in evidence E to the variable of interest D. The method is used to determine and organize the chains of reasoning from the variables with crucial findings to the variable of interest. The fundamental idea is to apply methods with different computational costs, starting with the simplest and fastest methods, and then step by step reducing the number of nodes, arcs and chains included in explanation.

First, INSITE uses structural information to determine the smallest set of nodes that are computationally related to variable of interest D given evidence E. *Computational relatedness* is the property that describes whether we need to know about the variable or not if we want to compute $P(D|E)$. Identification of computationally related nodes allows us to eliminate other nodes from the explanation. There are several algorithms to determine computational relatedness through analysis of the network structure and position of the evidence nodes E (Geiger, Verma, and J., 1989, 1990; Shachter, 1990; Baker and Boult, 1990). Criterion for computational relatedness is as follows:

Definition 14. *Node N_i is* computationally related *to D if and only if (1) N_i is a predecessor of D or predecessor of a member of evidence E, (2) N_i is connected to D by a path of nodes that are each computationally related to D, and (3) N_i is not d-separated (Definition 6) from D by members of E.*

As can be seen from Definition 14, elimination of computationally unrelated nodes can be done based on graphical criteria, regardless of probability distribution of the nodes involved.

Second, once computationally related nodes have been identified, direct chains between crucial evidence nodes and node of interest are identified using only computationally related nodes. Suermondt (1992) defines a direct chain between two nodes given evidence E and variable of interest D as follows:

Definition 15. *A* direct chain *between two nodes N_1 and N_k, given a variable of interest D and the evidence E, is defined as a sequence of distinct nodes $(N_1, \ldots, N_k)$ such that, (1) in the belief network, there exists an arc from N_i to N_{i+1} or from N_{i+1} to N_i for each $i \in \{1, \ldots, k-1\}$; and (2) every node in $N_1, \ldots, N_k$ is computationally related to D given E.*

Consider the example in Figure 10. All nodes in the Bayesian network in Figure 10 are computationally related (see Definition 14) to D, given evidence $E = \{E_1\}$. There are two chains from finding E_1 to the node of interest D: (E_1, N_2, N_3, D) and (E_1, N_2, N_4, D). Even though node N_1 is computationally related to D, it is not part of any direct chain from E_1 to D.

Before an algorithm starts to analyze direct chains, *nuisance nodes,* are removed from the network:

Definition 16. A *nuisance node*, given the evidence E and variable of interest D, is a node that is computationally related to D given $\boldsymbol{E}$ but that is not a part of any direct chain from any member of E to D.

For example, the node N_1 in the Bayesian network in Figure 10 is a nuisance node.

In the last part of qualitative analysis of the network, INSITE removes duplicate parts of chains between two knots. A knot is defined as follows:

Definition 17. A *knot* in a set S of direct chains from finding E_i to node D given a set of evidence E, is a node K_j such that (1) K_j is in every chain in S, and (2) $K_j \cup (E \setminus E_i)$ d-separates E_i from D.

Example 4. *The example in Figure 11 shows a small Bayesian network with a knot. There are two direct chains between nodes E_i and D: (E_i, N_1, K, D) and (E_i, N_2, K, D). Node K is a knot. Between E_i and K are two subchains:(E_i, N_1, K) and (E_i, N_2, K). Between nodes K and D there is only one subchain, (K, D), which is part of both of chains between E_i and D. Hence, by identification of the knot K, we can avoid redundant analysis of subchains between nodes K and D.*

Qualitative analysis of direct chains Graphical criteria select direct chains that are only potentially relevant to the inference result. Some direct chains may not be relevant at all. However, graphical criteria may not eliminate such direct chains from the explanation. Whether the direct chain $C\,(E_i, \ldots, D)$ should be included in the explanation is determined by the probabilities in the network. The INSITE method eliminates irrelevant nodes using *the strengths of direct chain* measure. *Strength of chain* is determined by the INSITE method by performing the following analysis:

1. Screening of chains based on a comparison of prior to posterior marginal distributions of nodes in the chain.
2. Determination of the effect of chains on the variable of interest.
3. Analysis of local effects of arcs within a single chain.

First, chains with a weak influence on the inference result are removed from the explanation. The term *proctored node* is used in the screening rule to select which nodes will be included in analysis of direct chains. A proctored node is defined as follows:

Definition 18. *A* proctored node *in a direct chain is the node that (1) is adjacent to two of its parents within a direct chain, and (2) is an evidence node or has at least one successor that is an evidence node.*

A proctored node facilitates propagation of the evidence among its parents, regardless of how much its own probability changes due to evidence. Figure 12 shows an example of proctored nodes.

Comparison of prior and posterior marginal probability distribution for the nodes in chain C is a computationally inexpensive way to determine whether the chain C affects the inference results. INSITE uses the following screening rule:

For all chains C from E_i to D, eliminate chain C if there is an N_j such that (1) N_j is not proctored in the chain C, and (2) $H\left(P\left(N_j|E\right);P\left(N_j\right)\right) \leq \theta$.

The chain screening method is based on cost of change between prior probability distribution and posterior probability distribution of the non-proctored node N_j in the chain C. Cost of change for the non-proctored node N_j given evidence E is compared to the threshold value θ to decide if chain C is to be removed from the explanation. If the cost of change is lower than the threshold value θ for some node from chain C, chain C is removed from the explanation. The screening rule eliminates chains with a low impact of evidence on the distribution of the non-proctored nodes in chain C. Cost of change between the prior probability distribution and the posterior probability distribution of the non-proctored node N_j in any chain C is the measure which is compared to the threshold value θ to decide which direct chain is to be removed from the explanation.

Arc removal Even after chain screening, irrelevant chains may remain. Another step used in INSITE to eliminate irrelevant chains is arc removal. Removal of arc XY may affect every chain that contains the arc XY and so allows evaluation of the combined role of the arc on transmission of evidence. Again, cost function H is used to measure the effect of removal of

an arc XY on evidence transmission. If the difference between the prior probability $P(D)$ in the original network and the prior probability $P'(D)$ in the network with the removed arc is significant, it is difficult to interpret the effect of arc removal on transmission of the evidence to the node of interest D. However, if cost of change in prior distribution due to arc removal $H(P(D), P'(D))$ is not significant and the cost of change in posterior probability due to arc removal is significant (i.e. $H(P(D|E); P'(D|E)) > \theta$), the arc is necessary for evidence transmission from E to D, otherwise arc will be removed.

Conflict analysis Next, INSITE performs conflict analysis for chains, which is similar to conflict analysis of evidence. It identifies if the chains of reasoning contribute to the inference result or conflict with the inference result. This method is based on comparison of size and direction of change. Chains of reasoning involving an $arc(X, Y)$ are consistent with the overall inference result if

$$H(P(D \mid E); P'(D \mid E)) \leq H(P(D \mid E); P(D))$$

and

$$Dir(P(D \mid E); P'(D \mid E)) = Dir(P(D \mid E); P(D)) ,$$

where P' is posterior probability if the $arc(X, Y)$ is removed.

Local effect of the arcs in the chain. Analysis of the local effect of the arc between two nodes X and Y in the chain C on the nodes X and Y is based on the measurement of the cost of change of posterior probabilities $P(Y|E)$ and $P(X|E)$ due to removal of the arc between nodes X and Y and the direction vectors $Dir(P(X|E); P'(X|E))$ and $Dir(P(Y|E); P'(Y|E))$, where P' is posterior probability if the $arc(X, Y)$ is removed. Details of this method can be found in Suermondt (1992).

Adaptation, communication and complexity Explanations are provided in the form of graphics and text. Probabilities are expressed numerically and verbally. INSITE does not have a user model but its explanation level of detail can be adapted to the user's needs. One disadvantage of the INSITE method is its computational complexity, which is exponential in the number of nodes and arcs in the worst case. One source of complexity is the complexity of inference itself in Bayesian networks. Another source is the combinatorial complexity when searching for a set of important findings and chains of reasoning. Despite the computational complexity, INSITE can run in a feasible amount of time on the 37-node ALARM network.

Evaluation of explanation INSITE was evaluated with the ALARM belief network, which was designed to help anesthesiologists with interpretation of monitored data. Human subjects, clinicians, were asked to evaluate patient cases, first without using any aid and later using only ALARM's advice or using ALARM's advice together with explanation provided by INSITE. INSITE was evaluated in five categories: diagnoses, actions, findings, confidence and opinions.

In the category "diagnoses", a differential diagnoses obtained after ALARM's consultation both with and without explanation were compared. Although not all results were not statistically significant, the results suggested that INSITE's explanation can potentially improve diagnostic performance of users. Among the results that were statistically significant was increase in number of incorrectly diagnosed cases after consulting ALARM without INSITE's explanation. Another statistically significant result was the smaller increase in number of new incorrect diagnoses if users consulted ALARM with INSITE's explanation in comparison to the increase of new incorrect diagnoses if users consulted ALARM without INSITE's explanation.

In the category "actions", actions written down by subjects as a response to question: what would you do next? Fewer actions were suggested and number of incorrect actions was lower if cases were evaluated with INSITE's explanation. However, this results were not statistically significant.

In the category "findings", findings identified by subjects as the most influential in the diagnosis were analyzed. The number of findings obtained if cases were evaluated with

INSITE's explanation was smaller than if the cases were identified without the explanation. However, the differences between findings identified with the explanation and without the explanation were not statistically significant.

In the category "confidence", confidence of the subjects in their judgment of cases measured through actions they selected , was compared for cases evaluated with INSITE's explanation and for cases evaluated without use of INSITE's explanation. Domain experts specified relative confidence of actions, taking into account the difficulty and the invasiveness of actions. Results showed increased confidence of subjects for cases which were evaluated with INSITE's explanation. These results were statistically significant.

Finally, in the category "opinions", subjects were asked to rate the following features on four-point scale (1=useless; 4=helpful): posterior probabilities, distribution of findings, effect of evidence on intermediate nodes, identification of diagnoses with greatest change in probability, identification of key evidence, conflict analysis, chain of reasoning (relationship between evidence and conclusion). Subjects then answered the following subjective questions for on each of the cases: the helpfulness of the computer's reasoning, the scope of ALARMS's model, the clarity of the computer's presentation. Each of the questions was answered using a rating on a seven-point scale (1=too simplistic, 4=captures essence, 7=too complex). Some statistically significant results were obtained. The number of cases when ALARM's model was ranked more than "captures essence" (>4) and less than "captures essence" (<4) was higher for cases with INSITE's explanation, while number of cases when ALARM's model was ranked "captures essence" and less than "captures essence" was higher for cases with INSITE's explanation. Out of the seven features, the one with the highest ranking was "automatic identification of diagnosis with greatest change in probability". The lowest ranked features were "effect of evidence on intermediate nodes", "definition of findings", "posterior probabilities".

3.3.5 Graphical display of the weight of evidence

Madigan et al. (1997) proposed method for visualizing probabilistic "evidence flows" in Bayesian networks. The method provides explanation of inference and also provides a test

selection facility. Explanation is displayed in the form of undirected graphs. The authors decided to use undirected graphs because they observed that explanation in directed graphs can often be counter-intuitive. The graphical–belief explanation methodology requires a unique path between the evidence and each node of interest (tree structure). The general Bayesian network must be transformed to a tree to satisfy this requirement. In order to apply this method to a wider class of Bayesian networks, Madigan and Mosurski (1990) designed the SAHR algorithm, which converts a Berge network into a network in which there is a unique path between each evidence node and node of interest. Madigan et al. (1997) also suggested an interactive method that uses clustering to convert a general Bayesian network into a Berge network, which could then be transformed into a tree structure using the SAHR algorithm (see Madigan et al. (1997) for details). While Suermondt (1992) and Haddawy et al. (1994) use measure of influence of evidence on a selected node to eliminate irrelevant chains and simplify the explanation, Madigan et al. (1997) use a SAHR algorithm (Madigan and Mosurski, 1990) to transform a network so that there is a unique path between the finding and the node of interest. The SAHR algorithm also marks important evidence. The explanation method uses *weight of evidence (WOE)* to measure influence of findings on the nodes in a Bayesian network:

$$W\left(D:E\right)=100\,log_{10}\frac{P\left(E|D\right)}{P\left(E|\neg D\right)}, \tag{3.2}$$

where E is evidence and D is the node of interest. Weight of evidence has a similar purpose to that of the Suermondt's cost of evidence (1992) and is used as a measure of quality of explanation. WOE determines strength and direction of influence of evidence on the selected node.

First, Madigan et al. (1997) used WOE to calculate the relative impact of each finding in the evidence on the node of interest. If evidence consists of n findings $E_1,\ldots,E_i,\ldots,E_n$, the weight of evidence E_i is

$$W\left(D:E_i\right)=100\,log_{10}\frac{P\left(E_i|D,E_1,\ldots,E_{i-1}\right)}{P\left(E_i|\neg D,E_1,\ldots,E_{i-1}\right)}.$$

WOE depends on the ordering of the findings in the evidence set chosen by the user. The advantage of WOE is that WOE, e.g. $W(X_1 = 1|X_n = 1)$, can be calculated recursively (see Madigan et al. (1997) for details) and therefore faster.

Second, the explanation method calculates *relevant outgoing weigh of evidence* in order to quantify potential outgoing WOE available, if the state of the intermediate variable were known. In order to demonstrate, assume a simple evidence chain as is in Figure 13 with the binary parameters and parameters in Table 2. Assume that we observed evidence for node $X5 = 1$. *Relevant outgoing weight of evidence* for node $X4$ given the evidence $X5 = 1$ is defined as

$$W_{rel:X5=1}(X3 = 1 : X2) = \begin{cases} W(X3 = 1 : X4 = 1) \text{ if } W(X4 = 1 : X5 = 1) > 0 \\ W(X3 = 1 : X4 = 0) \text{ if } W(X4 = 1 : X5 = 1) \leq 0 \end{cases}.$$

The width of the channels between the nodes in Figure 13 represents relevant outgoing weights of evidence. In our case $W(X4 = 1 : X5 = 1)$ =60.206 , therefore

$$W_{rel:X5=1}(X3 = 1 : X2) = W(X3 = 1 : X4 = 1) = 184.51.$$

Since there is no intermediate node between X5 and X4,

$$W_{rel:X5=1}(X4 = 1 : X5) = W(X4 = 1 : X5 = 1) = 60.206.$$

As can be seen from Figure 13, the potential WOE (relevant outgoing weights of evidence) between $X4$ and $X3$ is larger than the potential WOE between X5 and X4. The width of the interior bar represents the actual WOE calculated using Equation 3.2,

$$W(X4 = 1 : X5 = 1) = 60.206$$

for nodes X5 and X4 and $W(X3 = 1 : X5 = 1) = -21.2608$. Since evidence X5=1 supports a negative state of X3 (X3=0), and the actual weight of evidence $W(X3 = 1 : X5 = 1) < 0$ the inner band between X4 and X3 in Figure 13 is colored red. On the other hand evidence X5=1 supports a positive state of X4 (X4=1); therefore $W(X4 = 1 : X5 = 1) > 0$ and the inner band between nodes X5 and X4 in Figure 13 is colored blue. Comparison of actual WOE to potential WOE and incoming and outgoing WOE illustrates how restricted potential evidence flow is. Actual and potential weight of evidence for intermediate nodes is displayed graphically as a part of explanation (Figure 13).

3.3.6 Alternative measure of explanation quality

Chajewska and Draper (1998) criticize Suermondt's (1992) choice of cross-entropy as the measure of quality of explanation. They point out that cross-entropy is not a distance measure, ϱ, since it does not satisfy the requirements for a distance measure which are: symmetry ($\varrho(P_1, P_2) = \varrho(P_2, P_1)$) and triangle inequality ($\varrho(P_1, P_2) + \varrho(P_2, P_3) \geq \varrho(P_1, P_3)$). Chajewska and Draper (1998) also note that Suermondt's method does not compare the distribution change due to evidence change to the prior distribution of the node of interest, which could lead to incorrect conclusions about quality of explanation. The symbols δ_A, δ_B, X, δ_X, P_X, and D have the same meaning as in Definition 8. According to Chajewska and Draper (1998) the quality of explanation $X{=}\{\delta_x, P_x\}$ is judged based on the closeness of probability of D, given the explanation X, $P(d \mid \delta_X)$, to the posterior probability $P(d \mid \delta_B)$. There may be an explanation $X' = \{\delta_{X'}, P_{X'}\}$, where the distances between of probabilities $Dist(P(d \mid \delta_X), P(d \mid \delta_B))$ and $Dist(P(d \mid \delta_{X'}), P(d \mid \delta_B))$ are equal and

$$Dist(P(d \mid \delta_X), P(d \mid \delta_A) < Dist(P(d \mid \delta_{X'}), P(d \mid \delta_A)))$$

and

$$Dist(P(d \mid \delta_X), P(d \mid \delta_A) < Dist(P(d \mid \delta_X), P(d \mid \delta_B))).$$

Chajewska and Draper (1998) argue that in such case, $X' = \{\delta_{X'}, P_{X'}\}$ is a better explanation than $X = \{\delta_X, P_X\}$ and that cross-entropy is not a good quality of explanation measure since does not take into account distance of prior probability. Chajewska and Draper (1998) propose a set of requirements which a measure of quality of explanation should satisfy. First, the measure should be based on cost function, with its value depending on the size of the explanation set. The idea is to have an exhaustive but simple explanation. Second, the cost function f should take as arguments probabilities $P'(d \mid \delta_X)$, $P(d \mid \delta_A)$, and $P(d \mid \delta_B)$ for each value of the node of interest $d_i \in D$. The cost function should have the following properties: (1) The function should be monotonically increasing for $0 \leq P'(d \mid \delta_X) \leq P(d \mid \delta_B)$ and monotonically decreasing (for some applications monotonically non-decreasing) for $P'(d \mid \delta_X) > P(d \mid \delta_B)$. This implies that a measure achieves

its maximum if the explanation X is equal to the second instantiation δ_B:

$$\forall \delta_X \, f\left(P\left(d \mid \delta_X\right), P\left(d \mid \delta_A\right), P\left(d \mid \delta_B\right)\right) \leq$$
$$f\left(P\left(d \mid \delta_B\right), P\left(d \mid \delta_A\right), P\left(d \mid \delta_B\right)\right),$$

(2) An explanation that does not change its prediction from what is given by the first instantiation δ_A is worthless and it's quality measure should be 0, i.e.:

$$f\left(P\left(d \mid \delta_A\right), P\left(d \mid \delta_A\right), P\left(d \mid \delta_B\right)\right) = 0\,.$$

Chajewska and Draper (1998) propose two cost functions that satisfy the requirements specified above: ratio of relative differences and ratio of absolute differences. Ratio of relative differences is given by:

$$f_1(d) = 1 - \frac{\left|log\frac{P(d|\delta_B)}{P(d|\delta_X)}\right|}{\left|log\frac{P(d|\delta_B)}{P(d|\delta_A)}\right|} \tag{3.3}$$

and ratio of absolute differences by:

$$f_1(d) = 1 - \frac{P\left(d \mid \delta_B\right) - P\left(d \mid \delta_X\right)}{P\left(d \mid \delta_B\right) - P\left(d \mid \delta_A\right)}\,. \tag{3.4}$$

However, Chajewska and Draper (1998)did not empirically evaluated these new measures with human users.

3.3.7 An efficient explanation algorithm on polytrees

Similar to other explanation methods (Suermondt, 1992; Haddawy et al., 1994; Madigan et al., 1997), the method proposed by Chajewska and Draper (1998) explains which evidence causes a surprising change in the probability distribution of the node of interest and which nodes are relevant in transmitting the influence and the set of relevant paths P_x. However, different causal mechanisms that involve different evidence sets may cause the same change in probability distribution. Therefore, it is important to preserve the original causal mechanism. In order to assure that this is done, Chajewska and Draper (1998) proposed an algorithm which calculates the quality of explanation not only with respect to the node of interest,

but also with respect to intermediate nodes. Although this algorithm does not guarantee that the explanation produced by the algorithm will have the smallest explanation set for a given accuracy threshold, it guarantees that all the nodes in the explanation set will be bounded by an accuracy threshold. They propose using the quality of explanation measures described in Section 3.3.6. The complexity of the algorithm is $O\left(NV^B\right)$, where N is number of nodes, V is number values per node, and B is the branching factor. *The branching factor* is the number of children of a node. The advantage of this algorithm is that complexity is linear in the number of nodes for polytrees. However, since most real networks are not polytrees, in order to apply the explanation method to any Bayesian network it is necessary to transform networks with loops into polytrees. While this is possible, it imposes additional computational costs. Even though the number of nodes decreases after transformation, the number of values per node and the branching factor may increase substantially. In an extreme case, the complexity of the algorithm is $O\left(V^{N|\Delta|}\right)$, where Δ is the set of instantiated nodes and $|\Delta|$ is the number of instantiated nodes.

This method was designed for a causal Bayesian network and does not include user model, adaptation, or design of user-system interaction. The method was not empirically evaluated with or without users.

3.3.8 BANTER

The BANTER system was developed by Haddawy et al. (1994) as a shell to tutor the user in the evaluation of hypotheses and the selection of an optimal diagnostic procedure. The purpose of the BANTER system is to provide information about the Bayesian network in an intelligible form. One of the features provided by BANTER is explanation of reasoning. The explanation method used by BANTER is based on Suermondt's INSITE method. Haddawy et al. (1994) uses the term *influential findings* in the same way as Suermondt (1992) uses *relevant findings*. Like INSITE, BANTER also does explanation of reasoning in two steps: in the first step, BANTER identifies the relevant findings, and in the second step BANTER finds relevant chains of reasoning.

Relevant findings While INSITE uses cross-entropy to measure quality of explanation, BANTER uses influence and impact measures to identify relevant findings. Impact and influence measures are based on the information measure $I(d_j|E_i = e_i) = log\frac{P(d_j|E_i=e_i)}{P(d_j)}$, which measures how much more information is provided by the evidence $E_i = e_i$ about the event $D = d_j$. If $I(d_j|E_i = e_i) < 0$, evidence e_i causes a decrease in probability of d_j, and if $I(d_j|E_i = e_i) > 0$, evidence e_i causes an increase in the probability of d_j. Like INSITE, BANTER too allows identification of conflicting evidence. The product of $I(d_j|E_i = e_i) \cdot I(d_j|E = e)$ is used in BANTER to identify conflicting evidence. If

$$I(d_j|E_i = e_i) \cdot I(d_j|E = e) \ll 0$$

(symbol "$\ll$" means *much less than*), the change in probability caused by a single piece of evidence e_i *strongly disagrees* with the changes caused by evidence E. If $I(d_j|E_i = e_i) \cdot I(d_j|E = e) \gg 0$, the change in probability caused by a single piece of evidence e_i *strongly agrees* with the change caused by evidence E. The overall effect of the piece of evidence E_i is measured using an influence measure that is defined by the following formula:

$$\text{iinfluence}(D; E; E_i) = \sum_{d_j \in D} I(d_j; E) \cdot I(d_j; E_i).$$

The overall effect of a piece of evidence on the node of interest D without regard to direction is determined by a BANTER impact that is calculated using the following formula:

$$impact(D; E_i) = \sum_{d_j \in D} |I(d_j; E_i)|.$$

Both the INSITE and the BANTER system select relevant findings by comparing a selected quality of explanation measure to a predefined parameter θ. INSITE and BANTER implemented two different versions of a one-way analysis of evidence. INSITE searches for a set of relevant findings backwards, starting with all findings and eliminating irrelevant findings one by one. BANTER, on the other hand, does forward selection, starting with an empty set of relevant findings and adding relevant findings one by one to the set. Each of these approaches has pluses and minuses, which are discussed in Haddawy et al. (1994).

Chains of reasoning The method for finding direct chains in BANTER has the same purpose as the one in INSITE – to remove all barren nodes and nodes d-separated from the node of interest.

Unlike INSITE, which identifies direct chains in two steps (Section 3.3.4), BANTER simplifies the search for direct chains, combining the step for generating direct chains with the step for identifying d-separated nodes to perform identification into only one step.

However, like INSITE, BANTER too selects the direct chains with the highest strengths. As in measuring quality of explanation, BANTER uses impact as the measure to determine the strength of the direct chain, while INSITE uses cross-entropy. BANTER calculates $impact(N; E_i)$ for every node in a direct chain. The strength of a chain is determined by a minimum of impact values for the nodes in the chain.

BANTER is designed for causal Bayesian networks and generates only verbal explanations with probabilities being expressed quantitatively.

3.3.9 B2

B2 is a tutoring shell for a Bayesian network which was designed to allow medical students to practice medical decision-making (McRoy et al., 1996). B2 was a proposed extension of the BANTER system (Haddawy et al., 1994). The purpose of B2 is to improve the usability and usefulness of BANTER. B2 adds three major improvements to the original BANTER system. First, knowledge about the structure of a medical domain is added. Second, a discourse model for arguing about the content and structure of interaction is added. This allows elimination of irrelevant information. The last addition to B2 is natural language and graphical user interface gestures that allow the user to have better interaction with the system. The objective of a B2 explanation is comprehension of reasoning. Probabilities are expressed quantitatively. B2 allows hypothetical reasoning as well. There is no user adaptation in B2. The user-system interaction is facilitated by windows and questions are expressed in natural language.

3.3.10 Other explanation methods

The method developed by Yap et al. (2008) provides explanation in terms of influential nodes that are found in the Markov blanket of the target node. The explanation is based on restructuring part of the network containing target and influential nodes by reversing arcs such that influential nodes become ancestors of the target node. Then they generate an explanation in the following form: "BN predicts *<name of the target variable>* is *<the most likely state of the target variable>* with probability *<probability>* because *<name of some variable x> is <the most likely state of the variable x>* with probability *<probability >, because ... .*"

ExplainD developed by Poulin et al. (2006) provides visual explanation of evidence in additive classifiers e.g., naive Bayes, linear SVM, logistic linear models. The method defines 'weights of the classifier' which in case of naive Bayes corresponds to Good's 'weights of evidence' (Good (1985); Madigan et al. (1997)), which are used to measure the importance of an evidence.

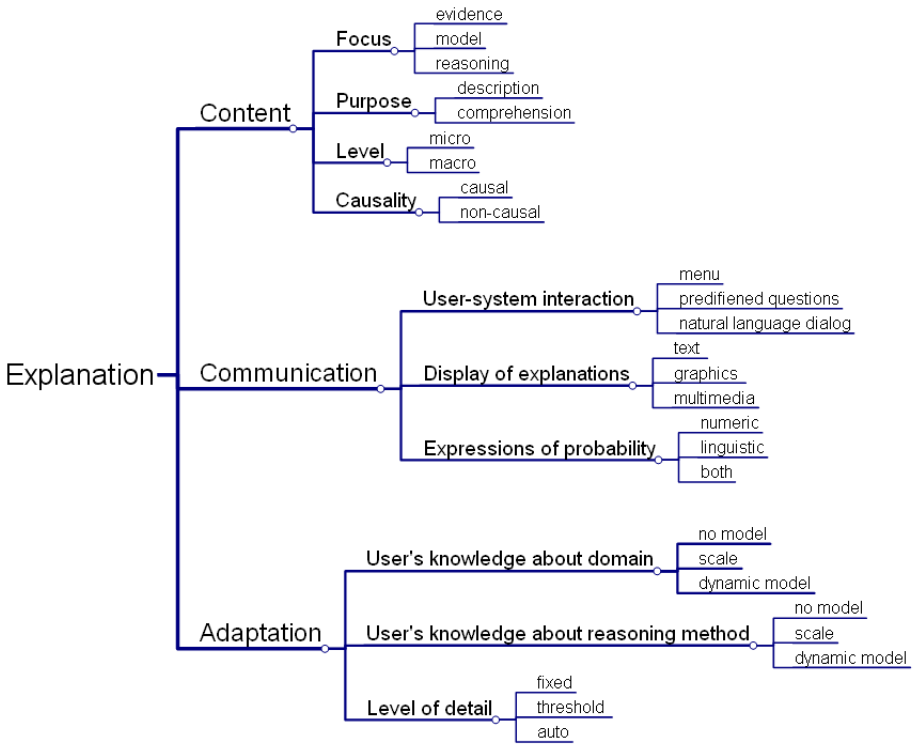


Figure 4: Explanation categorization adapted from Lacave and Diez (2002).

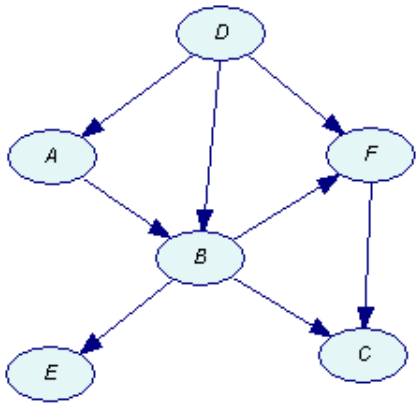


Figure 5: Bayesian network for Example 3

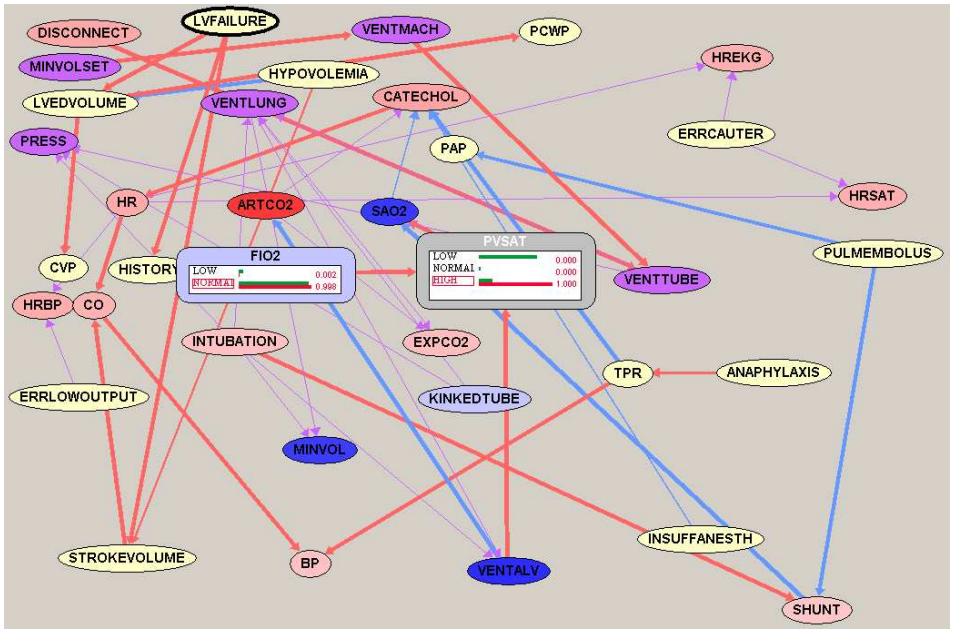


Figure 6: Graphical explanation in Elvira.

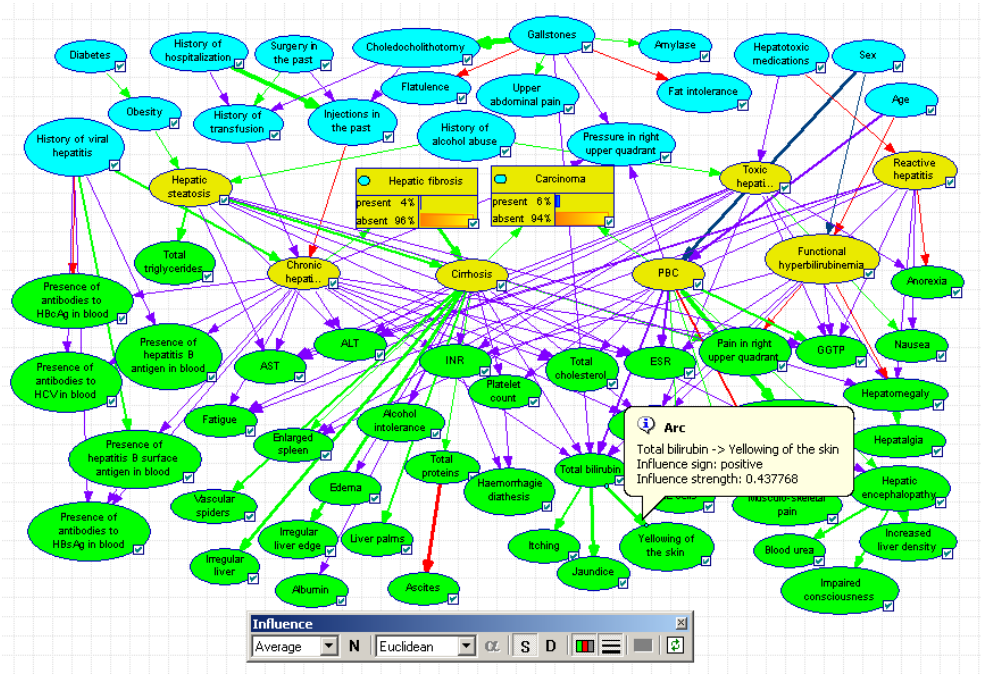


Figure 7: Graphical explanation in GeNIe.

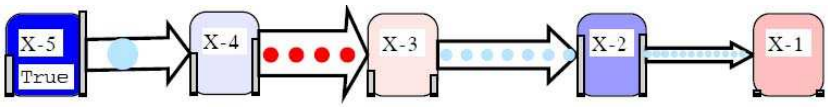


Figure 8: Alternative way to display weight of evidence in simple evidence chain using Petri display (Madigan et al., 1997). The width of the circles represents the actual strength of evidence for the next node in the chain. The color is blue for evidence supporting the "positive" state and red for evidence supporting the "negative" state. The balls move to follow the direction of evidence.

Evidence FOR Peptic Ulcer		*Evidence AGAINST Peptic Ulcer*	
Abdominal pain	(+9)	Length of history less than 1 year	(−75)
Episodic	(+19)	No previous operation for ulcer	(−5)
Relieved by food	(+44)	No seasonal effect on pain	(−9)
Occasionally woken at night and relieved by snack	(+25)	No waterbrash	(−29)
Epigastric	(+28)		
Point at site of pain with fingers	(+19)		
Family history of ulcer	(+39)		
Smoker	(+41)		
Vomits, then eats within 3 hours	(+54)		
	+278		−118
Balance of evidence	+160	(Total evidence 396: conflict ratio = 2.5)	
Initial score	−84	(corresponding to prevalence of 30%)	
Final score	+76	= 68% chance of peptic ulcer	

Figure 9: Evidence balance sheet which uses weight of evidence in centibans to quantify the contribution of findings in evidence to posterior probability of peptic ulcer (Spiegelhalter and Knill-Jones, 1984).

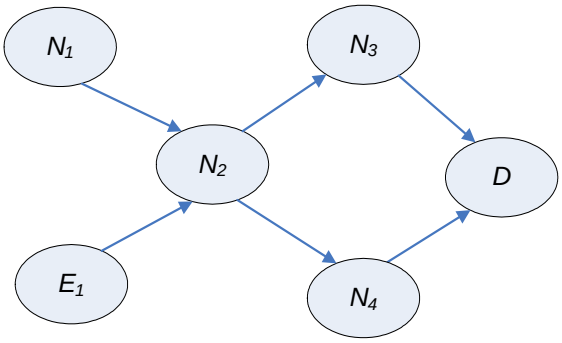


Figure 10: A small Bayesian network. All nodes are computationally related. There are two chains between nodes E_1 and D: (E_1, N_2, N_3, D) and (E_1, N_2, N_4, D). Node N_1 is not part of direct chain from E_1 to N_1.

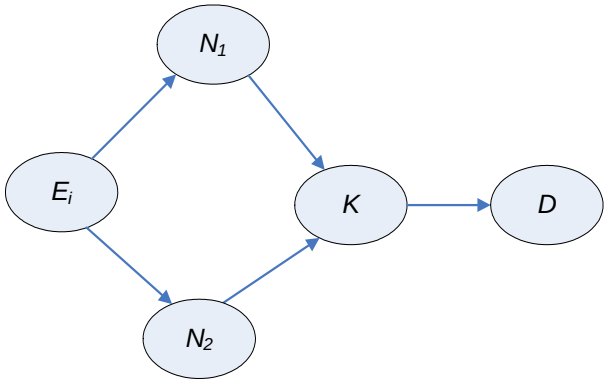


Figure 11: Knot. Small Bayesian network with two direct chains from E_i to D: (E_i, N_1, K, D) and (E_i, N_2, K, D). Node K is a knot.

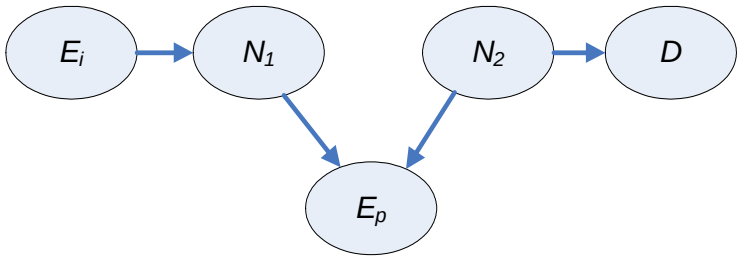


Figure 12: Proctored node. Nodes E_i and E_p are evidence nodes. There is a direct chain (E_i, N_1, E_p, N_2, D) from E_i to D. E_p is proctored since its neighbors, N_1 and N_2, are parents of E_p.

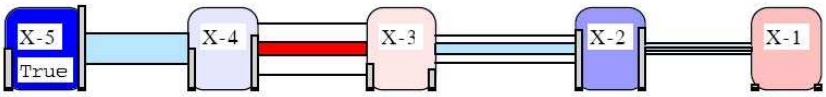


Figure 13: Simple evidence chain. The width of the channel represents the potential of evidence available (if the values of variables were known exactly). The width of the internal band represents the actual weight of evidence for the next node in the chain. A blue color denotes support of evidence towards a positive state and red denotes support of evidence towards a negative state.

Table 2: Parametrization for evidence chain in Figure 13

$Pr(X5 = 1\|X4 = 1) = 0.80$	Pr(X4=1|X3=1)=0.35	Pr(X3=1|X2=1)=0.25	Pr(X2=1|X1=1)=0.08
$Pr(X5 = 1\|X4 = 0) = 0.20$	Pr(X4=1|X3=0)=0.50	Pr(X4=1|X3=0)=0.50	Pr(X2=1|X1=0)=0.25

4.0 EXPLAINING INFERENCE ON A POPULATION OF INDEPENDENT AGENTS

4.1 EXPLANATION IN AGENT-BASED POPULATION BNS

4.1.1 Agent-based population BNs

An agent-based population BN represents all agents in a population individually (see Figure 14). An agent-based BN consists of two main parts: a subnetwork representing the whole population (common part) and a subnetwork that represents agents in the population individually (population part), with each agent represented by its own subnetwork (Figure 15). The nodes that belong to the population part of the BN and are connected with the common part of the BN I call *interface nodes*. The size of the population can be very large. For example, a Bayesian model for biosurveillance, PANDA-CDCA (Cooper et al., 2006), which I describe in detail in Section 4.1.2, was tested for a population as large as 423,000 individuals. Individuals in PANDA-CDCA represent agents (Figure 15). In principle, there can be direct interactions between agents in agent-based population networks (Figure 16); however, the network in Figure 14 assumes that agents do not interact with each other, hence there are no directed arcs connecting variables in different agents' subnetworks.

4.1.2 PANDA-CDCA

PANDA-CDCA (Figure 15) is an agent-based Bayesian network for diagnosing outbreaks of CDC Category A diseases, namely anthrax, smallpox, tularemia, botulism and hemorrhagic fever, as well as several additional diseases, such as influenza, cryptosporidiosis, hepatitis A, and asthma. PANDA-CDCA uses emergency department chief complaints to diagnose

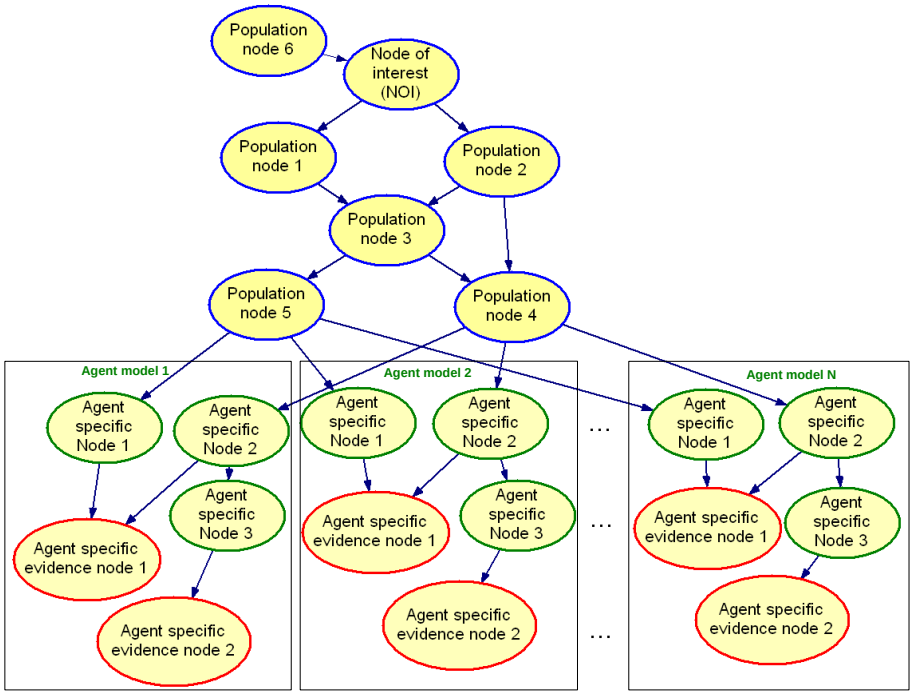


Figure 14: Agent-based Bayesian networks without interaction between agents in population. (Repeated appearance for reader's convenience)

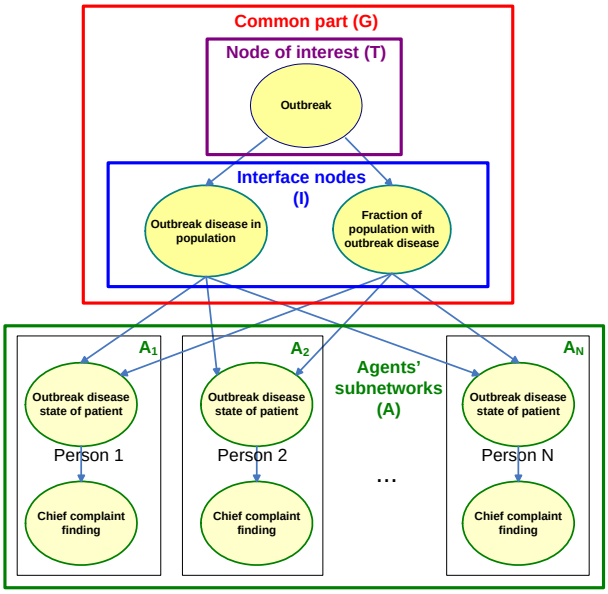


Figure 15: Agent-based network example. PANDA-CDCA (Cooper et al., 2006).

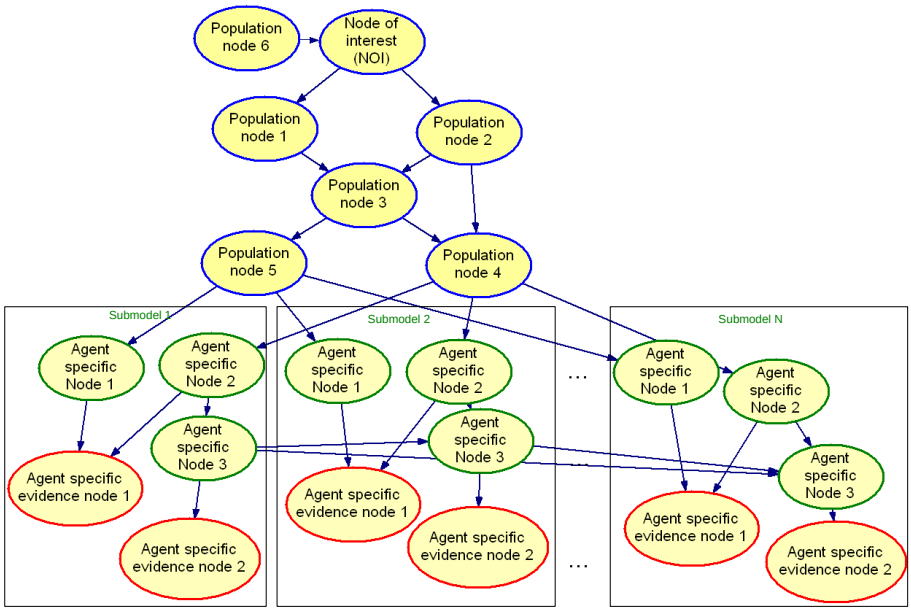


Figure 16: Agent-based Bayesian networks with interaction between agents in population . (Repeated appearance for reader's convenience)

disease outbreaks. The model takes as input emergency department (ED) chief complaints observed during the previous 24 hours, and every hour outputs the posterior probability of the modeled diseases. The model consists of a *Common part* (**G**) made up of nodes that represent features common to the whole population, and a *Population part* (**A**), consisting of subnetworks $\boldsymbol{A} = \{A_1, \ldots, A_n\}$ of all n individuals in the population, where A_i is the subnetwork of the i^{th} individual. The nodes in the *Common part* of the network are *Outbreak*, *Outbreak disease in population* and *Fraction of population with outbreak disease*. I will refer to the node *Outbreak disease in population* simply as *Outbreak disease* and to the node *Fraction of population with outbreak disease* as *Fraction of population*. The *Outbreak* node represents the presence or absence of an outbreak. The *Outbreak disease* node represents the explicitly modeled diseases listed above. *Fraction of population* represents the fraction of the population that has an outbreak disease and has come to the ED within the previous 24 hours. The nodes *Fraction of population* and *Outbreak disease* interface between the other *Common part* node (*Outbreak* node) and the *Population part* of the BN (**A**). PANDA-CDCA

assumes conditional independence of agents (Figure 15). This assumption is reasonable for non-contagious disease outbreaks in the population when we condition on the important factors that cause individuals to contract disease, such as the release of biological material (e.g., an aerosol of anthrax spores). PANDA-CDCA also models contagious diseases such as smallpox, where conditional independence does not strictly hold. The naive Bayes classification model (Domingos and Pazzani, 1997) also assumes a closely related form of conditional independence. Such models have been shown to perform classification remarkably well, even in domains where the conditional independence assumption is violated. PANDA-CDCA is built on the premise that disease-outbreak classification will be performed well even in modeling contagious diseases, where assumption of conditional independence is violated. In any case, my dissertation is only concerned with explaining PANDA-CDCA-like models in which conditional independence is a reasonable assumption, as would be the case in modeling non-contagious diseases in outbreak detection.

In PANDA-CDCA all individuals (agents) are represented as identical subnetworks. An agent's subnetwork consists of the nodes *Outbreak disease state of patient* and *Chief complaint finding*. The node *Outbreak disease state of patient* represents diseases that each person can have according to the model. The CDC category A diseases, plus influenza, cryptosporidiosis, hepatitis A, and asthma, are modeled explicitly; any other disease which the patient may have is represented by the state "other", meaning some other disease. PANDA-CDCA has been tested using semisynthetic and real data with encouraging results (Cooper et al., 2006), although additional evaluation is ongoing.

In an experimental evaluation of the *hierarchical explanation method* (HEM), I will use a simple variation of the PANDA-CDCA model so that the evaluation will be independent of whether study participants with knowledge of biosurveillance are available. The simpler Simple Biosurveillance Network (SBN) is described in Section 5.2.

4.1.3 How hierarchical explanation relates to existing explanation methods

The main aim of explanation in probabilistic systems is to explain how and why a particular posterior distribution of nodes of interest was obtained as an effect of the observed evidence

applied to the model. Among explanation methods that include an analysis of evidence are those developed by Suermondt (1992), Haddawy et al. (1994), and Chajewska and Draper (1998). All of these methods follow to some degree the explanation framework developed by Suermondt (1992). Since population BNs can contain hundreds or thousands of agent-specific subnetworks and at least the same number of findings, it would be impractical and even unfeasible to use the previously developed explanation methods. One reason is that methods for exact analysis of evidence require a complete search over all possible subsets of observed findings, where the candidate evidence subsets are scored based on ability to replicate the inference results obtained with the complete set. As mentioned earlier, some explanation methods use approximations in order to decrease computational complexity. Suermondt (1992) proposed a heuristic approach with a computational complexity that is linear in the number of findings. He also designed a test of reliability of results obtained using his heuristic method. The disadvantage of Suermondt's method is that it ignores the possibility that similar changes in the probability distribution of the *node of interest (NOI)* due to an evidence subset can be caused by different causal mechanisms. It is possible that the selected evidence could change the posterior probabilities of the *intermediate nodes* (nodes on the path between evidence and *NOI*) substantially in comparison with the probabilities obtained with the original evidence, while keeping the posterior of *NOI* within the tolerated range. Thus the INSITE method does not necessarily provide the same causal explanation as one based on the complete evidence would. Chajewska and Draper (1998) proposed checking the change in posterior probability of all nodes between evidence and *NOI* in order to preserve the causal mechanism. The cost of this approach is a higher computational complexity for BNs with structures more complex than polytrees. However, while the algorithm is efficient for polytrees, it does not guarantee finding the smallest evidence set that satisfies the required quality of explanation.

The previously developed methods also do not take advantage of the modular structure in the BN in organizing their explanation. Analysis of arcs, proposed in previous research, also may be impractical in agent-based networks because with a larger population size, the number of nodes and arcs to be analyzed is proportionally larger. In this paragraph I will briefly introduce the main principles of the method proposed in this study. I will

use in the text the term *directed tree*, meaning a polytree (Definition 9) in which every node has at most one parent (Figure 17). The *Hierarchical explanation method* (HEM) is

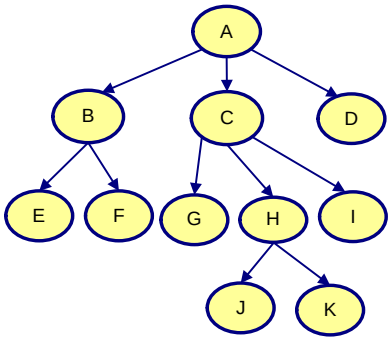


Figure 17: Example of directed tree

inspired by an explanation method developed by Spiegelhalter and Knill-Jones (1984) for explanation in directed trees based on a weight of evidence measure (WOE) (Good and Card, 1971; Good, 1977, 1985). Unlike most existing explanation methods, which begin obtaining the explanation with analysis of evidence, the HEM does analysis of evidence as the last step. The HEM produces an explanation in three steps, moving from the node of interest through interface nodes toward evidence nodes. HEM takes advantage of the possibility of naturally partitioning agent-based population networks due to independence of agents if interface nodes are observed. HEM takes advantage of groups of patients with similar evidence as well. Grouping agents with similar networks and evidence enables faster computation and simpler explanations. Once groups of agents with important evidence are identified, an existing explanation method (e.g., INSITE) can be applied to construct an explanation within an agent's network.

Section 4.2 presents a method that generates explanations of how findings of a population of agents influence population characteristics.

4.2 HIERARCHICAL EXPLANATION IN A BN WITH A POPULATION OF INDEPENDENT AGENTS

4.2.1 Introduction

Agent-based networks are useful in situations in which we want to learn something about a population based on information about the agents in the population. In disease-outbreak detection (biosurveillance), the agents are often people reporting their symptoms upon admission to the hospital. Another type of agent could be a sensor that periodically measures and reports information about air quality at a given location in a city. Another example of an agent-based BN is a BN that models eyewitnesses of traffic accidents as agents. Information about the accident is inferred from the BN and the testimonies of the eyewitnesses. An example of an agent-based BN in the military domain is the collection of intelligence from soldiers engaged in combat about the size and nature of an enemy force, in order to derive an estimate of the enemy's military capabilities. A further example of an agent-based BN is a BN for detecting the threat of terrorist attack using information from heterogeneous sources such as bank reports of suspicious financial transactions, satellite observations of unusual activity in the territory, and reported illegal purchase or theft of large volumes of explosives, biological, chemical, or nuclear weapons, or components for their production. All of these sources can be considered to be independent conditionally on the terrorist group and the location and type of attack.

The central idea of a HEM is to make use of the specific structure of agent-based BNs and the non-interaction among the agents. The structure of agent-based BNs enables us to select the agents that are most important for obtaining the inference results of interface nodes efficiently (Figure 15). Figure 18 provides the scheme of a HEM. The HEM builds up the explanation hierarchically using three levels. *Scores* are used to select the important states that are most relevant to the inference results for the observed evidence. HEM starts at the first level to select the most important (probable) states(s) of the first-level node(s) given the observed evidence. First level-nodes are *nodes of interest* (NOI), whose probability we want to explain. A *node of interest* (NOI) can be instantiated to states $noi_1, noi_2, \ldots, noi_{N_{\mathrm{NOI}}}$. A

score is used to quantify the importance of each instantiation for the explanation. Next, for the selected state(s), the most supportive state(s) of variables at the second level, interface nodes (I), are selected using the score described in Section 4.2.4. Interface node $\mathcal{I}$ can be instantiated to states $i_1, \ldots, i_{N_I}$. Finally, for the selected important state(s) of nodes at the second level, the most supportive agents, represented by submodels, are selected based on how much the available information about them supports the second-level configuration(s). In Section 4.2.2 I explain this concept of hierarchical explanation with an example.

4.2.2 Example of hierarchical explanation

I will provide an example of a HEM for PANDA-CDCA (Section 4.1.2). Details of hierarchical explanation, such as calculation of scores, clustering, and filtering of information are discussed in the following sections of Section 4.2. Figure 19 shows a schema of HEM applied to PANDA-CDCA. Observed evidence for PANDA-CDCA $\boldsymbol{e}$ consists of chief complaint findings (CCFs) extracted from the chief complaint string recorded for each patient arriving at the emergency department. The chief complaint findings considered by the model include difficulty of swallowing, slurred speech, and hemoptysis. In the case of PANDA-CDCA, the variable whose posterior probability we want to explain (node of interest) is *Outbreak*. First, all possible instantiations of the *Outbreak* node are scored using the posterior probability of the *Outbreak* node given the observed evidence. For the following example in Figure 19, the posterior probability of $Outbreak = true$ is 0.9999 and the posterior probability of $Outbreak = false$ is 0.0001. Assume we would like to know why the posterior of $Outbreak = true$ is so high. In the next step, HEM identifies the instantiations of the interface nodes. For simplicity, I will not include the node *Fraction of population with outbreak disease* in the explanation. All possible instantiations of the variable *Outbreak disease* are scored using the posterior of "outbreak disease" given evidence provided by CCF and $Outbreak = true$. Following the example in Figure 19, the top scored instantiation is $Outbreak\,disease = botulism$ ($score = 0.998$) and the second most highly scored instantiation is $Outbreak\,disease = plague$ (0.001). Assume we are most interested in the most important (highest scored) instantiation $Outbreak\,disease = botulism$ ($score = 0.998$). In

the last stage, HEM searches for the group of patients that provides the highest evidential support for the instantiation $Outbreak\,disease = botulism$. A score based on a conditional likelihood ratio is used to quantify the support that evidence observed for a given group of patients provides for the instantiation $Outbreak\,disease = botulism$. The likelihood ratio is given by the following equation:

$$L\,(OD = botulism : CCF = difficulty\,in\,\text{swallowing}) = \frac{P(CCF=difficulty\ in\ swallowing|OD=botulism)}{P(CCF=difficulty\ in\ swallowing|OD\neq botulism)}$$

The highest support (measured by the score) for $Outbreak\,disease = botulism$ given $Outbreak = true$ is provided by the group of 36 patients with the finding "difficulty in swallowing". The second highest support is provided by the group of patients with the finding "slurred speech".

4.2.3 Generation of explanation

After the scores discussed in Section 4.2.2 are calculated, HEM generates a default verbal explanation, which is presented to the user. This verbal explanation is complemented by a graph presenting an analysis of the evidence. In the example in Section 4.2.2, I used the version of explanation where only *Outbreak Disease* is explained as an interface node. The verbal explanation that HEM will generate for that example is "PANDA-CDCA detected an outbreak ($Outbreak = true$) with probability 0.9999. The most probable outbreak disease is botulism, with probability 0.998. Evidence that supports botulism as the outbreak disease is a group of 36 patients with the chief complaint of difficulty in swallowing. When 36 such patients come to the emergency department, the probability of botulism increases 22 times with respect to alternative outbreak diseases". This verbal explanation will be complemented by an analysis of the evidence as in Figure 20, which shows how different evidence groups increase or decrease posterior odds for the instantiation $OutbreakDisease = botulism$.

In the following sections I will describe in detail the calculation of the scores used in the explanation.

4.2.4 Explaining an instantiation of *NOI*

The first level of a HEM is represented by a node of interest (NOI). It is the posterior probability of this node that we would like to explain. In the case of a biosurveillance system, it is usually a variable that represents whether there is (or is not) an outbreak. In PANDA-CDCA, it is the outbreak node.

I assume that a *node of interest* (NOI) belongs to the common part of the network. In order to exploit the fact that agents do not directly interact with each other, we instantiate interface nodes in order to make agent subnetworks conditionally independent of each other.

HEM starts with an explanation of which states of the NOI are most and least probable given the observed evidence, where the score for state noi_j is given by the posterior probability of the j^{th} state configuration noi_j :

$$score(NOI = noi_j \mid \boldsymbol{E}) = P\left(NOI = noi_j | \boldsymbol{E}\right). \tag{4.1}$$

The NOI can be one node or a set of nodes that we are interested in. In general there could be many state configurations of NOI. However, including all configurations in the explanation is not optimal, since the explanation should include only relevant information. We are usually interested in explaining the most probable outcome. Explanation of the most probable outcome can be complemented by explanation of why some other outcome of interest is not the most probable given the evidence. I discuss methods for limiting the number of states to those most significant for explanation based on the score calculated according to Equation 4.1. Now, assume that we have selected a subset of configuration $\mathcal{C}^{+}\left(NOI\right)$ of the most important states for NOI that we would like to explain. In the case of PANDA-CDCA, the NOI is usually the binary variable *outbreak*. Which possible configuration of *Outbreak*, $Outbreak = yes$ or $Outbreak = no$, will be explained depends on the posterior probability of these states. I am using $\mathcal{C}^{+}$ to represent the selected set of high scoring instantiations and $\mathcal{C}^{-}$ to represent the selected set of low scoring instantiations. The symbol $\mathcal{C}$ represents a selected set of instantiations that can be either low scoring or high scoring.

In the next step, HEM explains why the score $S\left(NOI = c \mid \boldsymbol{e}\right)$ for configuration $c \in \mathcal{C}_j^{+}\left(NOI\right)$ is high. The score $S\left(NOI = c|\boldsymbol{e}\right)$ is explained in terms of joint configuration of

the interface nodes. There are three reasons why interface nodes are selected: (1) interface nodes d-separate agent subnetworks, (2) interface nodes d-separate agent subnetworks and the *NOI*, and (3) including interface nodes in the explanation provides additional insight into how evidence propagated from agents is aggregated and transmitted by intermediate nodes. The score that HEM uses to measure to what extent the joint configuration of the interface node $\mathcal{I} = i$ influences the score of the *NOI* is as follows :

$$S\left(\mathcal{I} = i \mid NOI = c, \boldsymbol{e}\right) = P\left(\mathcal{I} = i \mid NOI = c, \boldsymbol{e}\right) \tag{4.2}$$

By applying filtering and clustering of configurations for interface nodes $\mathcal{I}$ (Section 4.2.6), we can partition configurations in $\mathcal{C}^{+}$ into several subsets of high scoring configurations $\mathcal{C}_j^{+}\left(\mathcal{I}\right)$, ranked using the score $S\left(\mathcal{I} = i \mid NOI = c, \mathbf{e}\right)$ given by Equation 4.2. In a similar way, we can select a subset of configurations $\mathcal{C}_j^{-}\left(\mathcal{I} = i\right)$ that contradict the configuration $NOI = c$ for any $c \in C\left(NOI\right)$.

Once we have identified $\mathcal{C}\left(\mathcal{I}\right)$, we can start the next level of explanation and select evidence that is most supportive for the selected configuration of interface nodes. Each agent in the population is represented by an agent-specific subnetwork and evidence. In Section 4.2.5 I describe how HEM selects agents which support configuration $\mathcal{C}\left(\mathcal{I}\right)$. I will use $P\left(\neg i\right)$ to denote $P\left(\mathcal{I} \neq i\right)$ and $P\left(i\right)$ to denote of $P\left(\mathcal{I} = i\right)$ in the following text in order to simplify the notation.

4.2.5 Explaining an instantiation of the interface nodes using evidence about agents in the population

One way to select agents supporting $\mathcal{C}\left(\mathcal{I}\right)$ is to follow the approach of Madigan et al. (1997), which is described in Section 3.3.5 and Spiegelhalter and Knill-Jones (1984) described in Section 3.3.2, both of whom used weight of evidence (WOE) to construct their explanations. Assume we want to explain the influence of the findings in the evidence $\boldsymbol{e} = \{e_1, \ldots, e_N\}$ on the posterior probability of instantiation $\mathcal{I} = i_1$ with respect to $\mathcal{I} = i_2$, given some background evidence $\mathcal{B}$ whose explanatory power does not interest us. The explanation

methods mentioned above are based on the property of naive Bayes independence, under which we can decompose a ratio of posterior probabilities

$$O\left(i_1/i_2 \mid \boldsymbol{e}, \mathcal{B}\right) = \frac{P\left(i_1 \mid \boldsymbol{e}, \mathcal{B}\right)}{P\left(i_2 \mid \boldsymbol{e}, \mathcal{B}\right)},$$

into the product of the ratio of prior probabilities

$$O\left(i_1/i_2 \mid \mathcal{B}\right) = \frac{P\left(i_1 \mid \mathcal{B}\right)}{P\left(i_2 \mid \mathcal{B}\right)},$$

and likelihood ratios

$$LR\left(i_1/i_2 : e_j \mid \mathcal{B}\right) = \frac{P\left(e_j \mid i_1, \mathcal{B}\right)}{P\left(e_j | i_2, \mathcal{B}\right)} \text{ for } j \in \{1, \ldots, N\},$$

where e_j denotes one of the findings in the evidence $\boldsymbol{e}$. Decomposition of a ratio of posterior probabilities is shown in the following equation:

$$\frac{P\left(i_1 \mid \boldsymbol{e}, \mathcal{B}\right)}{P(i_2 \mid \boldsymbol{e}, B)} = \frac{P\left(i_1 \mid \mathcal{B}\right)}{P\left(i_2 \mid \mathcal{B}\right)} \prod_{j=1}^{N} \frac{P\left(e_j \mid i_1, \mathcal{B}\right)}{P\left(e_j \mid i_2, \mathcal{B}\right)}. \tag{4.3}$$

If I is a binary variable, $i_1 = i$ and $i_2 = \neg i$, the probability ratios $O\left(i_1/i_2|\mathcal{B}\right)$ and $O\left(i_1/i_2|\mathcal{B}, \boldsymbol{e}\right)$ become prior and posterior odds with respect to $\boldsymbol{e}$, given background knowledge $\mathcal{B}$ defined as:

$$O\left(i|\mathcal{B}\right) = \frac{P\left(i|\mathcal{B}\right)}{1 - P\left(i|\mathcal{B}\right)},$$

$$O\left(i|\mathcal{B}, \boldsymbol{e}\right) = \frac{P\left(i|\boldsymbol{e}, \mathcal{B}\right)}{1 - P\left(i|\boldsymbol{e}, \mathcal{B}\right)},$$

and the likelihood ratio is given by

$$LR\left(i : e_j|\mathcal{B}\right) = \frac{P\left(e_j|i, \mathcal{B}\right)}{P\left(e_j|\neg i, \mathcal{B}\right)}. \tag{4.4}$$

Log-transformation of Equation 4.3, substitution, $i_1 = i$, and $i_2 = \neg i$ allow Equation 4.3 to be expressed in terms of log-ratios and weight of evidence e:

$$log\left[\frac{P\left(i|\boldsymbol{e}, \mathcal{B}\right)}{P\left(\neg i|\mathcal{B}\right)}\right] = log\left[\frac{P\left(i|\mathcal{B}\right)}{P\left(\neg i|\mathcal{B}\right)}\right] + \sum_{j=1}^{N} log\left(\frac{P\left(e_j \mid i, \mathcal{B}\right)}{P\left(e_j \mid \neg i, \mathcal{B}\right)}\right),$$

where $WOE(i/\neg i : e)$ is given by the following equation (Good, 1977):

$$WOE(i/\neg i : e_j \mid \mathcal{B}) = log\left(\frac{P\left(e_j \mid i, \mathcal{B}\right)}{P\left(e_j \mid \neg i, \mathcal{B}\right)}.\right). \tag{4.5}$$

The relationship between WOE and the likelihood ratio is given by Equation 4.6.

$$WOE(i/\neg i : e_j \mid \mathcal{B}) = log\left(LR\left(i : e_j \mid \mathcal{B}\right)\right). \tag{4.6}$$

$WOE(i/\neg i : e \mid \mathcal{B})$ quantifies the contribution that e_j provides in favor of hypothesis $I = i$ as against hypothesis $I = \neg i$ given the background evidence $\mathcal{B}$. This feature has the advantage of allowing us to evaluate the contribution of each piece of evidence e_j independently using $WOE(i/\neg i : e_j \mid \mathcal{B})$. To simplify the notation I will use $WOE(i : e_j \mid \mathcal{B})$ to represent $WOE(i/\neg i : e_j \mid \mathcal{B})$ and $LR(i : e_j)$ to represent $LR(i/\neg i : e_j)$. $WOE(i : e_j)$ not only quantifies the support which the evidence provides for configuration i , but also allows us to distinguish which pieces of evidence are for and which are against configuration i . Based on the sign of $WOE(i : e_j)$, three main effects of evidence can be recognized:

$$WOE(i : e_j \mid \mathcal{B}) = \begin{cases} > 0 & \text{evidence } e_j \text{ supports instantiation } i \\ & \text{over instantiation } \neg i \text{ in the context of } \mathcal{B} \\ = 0 & \text{evidence } e_j \text{ supports instantiation } i \\ & \text{as much as instantiation } \neg i \text{ in the context of } \mathcal{B} \\ < 0 & \text{evidence } e_j \text{ supports instantiation } \neg i \\ & \text{over instantiation } i \text{ in the context of } \mathcal{B}. \end{cases} \tag{4.7}$$

Spiegelhalter and Knill-Jones (1984) used WOE to construct an evidence balance sheet with evidence for the disease on one side and evidence against the disease on the other side, sorted according to the score for each piece of evidence (Section 3.3.2). Tukey (Tukey's comments in Spiegelhalter and Knill-Jones 1984) proposed a fully graphical display of the contribution of an agent's evidence to the posterior probability resulting from the total population evidence (see Figure 21). I want to use a similar graphical display as a part of HEM.

I choose the *likelihood ratio* (LR) as a score for selection of supporting evidence and the generation of a verbal explanation for its intuitive interpretation. I choose WOE for a graphical analysis of the evidence for its additive property (Figure 20). The WOE is proportional to the monotonic (logarithmic) transformation of LR, as can be seen from Equation 4.5. The likelihood ratio $LR(i/\neg i : \boldsymbol{e}_j \mid \mathcal{B})$ quantifies the contribution that evidence $\boldsymbol{e}_j$ provides in favor of hypothesis $\mathcal{I} = i$ as against hypothesis $\mathcal{I} = \neg i$ in the context of evidence $\mathcal{B}$. I plan to use LR to measure evidential support. I assume that all pieces of evidence that we want to analyze in $\boldsymbol{e} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_j \ldots, \boldsymbol{e}_N\}$ come from agents in the population, where $\boldsymbol{e}_j$ is evidence that comes from the j^{th} agent. In general, evidence for agent j, $\boldsymbol{e}_j$ consists of L_j findings e_j^l, where $\boldsymbol{e}_j = \left\{e_j^1, e_j^2, \ldots, e_j^{L_j}\right\}$. Since the goal is to find agents whose evidence most supports the state configurations $\mathcal{C}(\mathcal{I})$ of the interface nodes, I want to select agents whose evidence $\boldsymbol{e}_j$ contributes most to the posterior probability $p(\mathcal{I} = \mathcal{C}(\mathcal{I}) | \boldsymbol{e}, \mathcal{C}(NOI))$ of configuration $\mathcal{I} = \mathcal{C}(\mathcal{I})$, relative to the posterior probability of all other possible instantiations $\mathcal{I} \neq \mathcal{C}(\mathcal{I})$ given evidence $\boldsymbol{e}$, $p(\mathcal{I} \neq \mathcal{C}(\mathcal{I}) \mid \mathcal{C}(NOI))$. In other words, we want to select agents with findings that contribute most to the posterior odds

$$O(I = \mathcal{C}(\mathcal{I}) \mid \boldsymbol{e}, \mathcal{C}(NOI)) = \frac{p(I = \mathcal{C}(\mathcal{I}) \mid \boldsymbol{e}, \mathcal{C}(NOI))}{p(I \neq \mathcal{C}(\mathcal{I}) \mid \boldsymbol{e}, \mathcal{C}(NOI))}, \tag{4.8}$$

where the general definition of odds for event x with the probability $p(x)$ is $O(x) = \frac{p(x)}{1-p(x)}$. If $\mathcal{I}$ is a binary variable, posterior odds can be decomposed into the product of prior odds given by Equation 4.9 and the conditional likelihood ratios given by Equation 4.10.

$$O(I = \mathcal{C}(\mathcal{I})) = \frac{p(I = \mathcal{C}(\mathcal{I}))}{1 - p(I = \mathcal{C}(\mathcal{I}))}. \tag{4.9}$$

$$LR(\mathcal{C}(\mathcal{I}) / \neg \mathcal{C}(\mathcal{I}) : \boldsymbol{e}_j) = \frac{p(\boldsymbol{e}_j \mid I = \mathcal{C}(\mathcal{I}))}{p(\boldsymbol{e}_j \mid I \neq \mathcal{C}(\mathcal{I}))}. \tag{4.10}$$

Factorization of the posterior odds is then given by following equation:

$$\frac{p(I = \mathcal{C}(\mathcal{I}) \mid \boldsymbol{e}, \mathcal{C}(NOI))}{p(I \neq \mathcal{C}(\mathcal{I}) \mid \boldsymbol{e}, \mathcal{C}(NOI))} = \frac{p(I = \mathcal{C}(\mathcal{I}))}{1 - p(I = \mathcal{C}(\mathcal{I}))} \prod_{j=1}^{N} \frac{p(\boldsymbol{e}_j \mid I = \mathcal{C}(\mathcal{I}))}{p(\boldsymbol{e}_j \mid I \neq \mathcal{C}(\mathcal{I}))}. \tag{4.11}$$

This allows us to measure the contribution of each agent's evidence, $\boldsymbol{e}_j$, to the posterior odds of instantiation $\mathcal{C}(\mathcal{I})$ of interface nodes given the instantiation $\mathcal{C}(NOI)$ of nodes of interest independently.

In general, however, $\mathcal{I}$ will be a set of multivalued variables. In such cases, either i , $\neg i$, or both represent multiple configurations of variables in $\mathcal{I}$. Let $\neg i$ represent multiple configurations of $\mathcal{I}$. In general, agents in a population are not conditionally independent of each other given $\mathcal{I} \neq i$, which can be formally expressed as the following inequality:

$$P\left(\{E_j\}_{j=1}^{k} \mid \neg i\right) \neq \prod_{j=1}^{k} P\left(E_j \mid \neg i\right). \tag{4.12}$$

We cannot then evaluate agents using their findings independently, and posterior odds must be decomposed into prior odds and conditional likelihood ratios as follows:

$$\begin{gathered}\frac{p(I=\mathcal{C}(\mathcal{I})|\boldsymbol{e},\mathcal{C}(NOI))}{p(I\neq\mathcal{C}(\mathcal{I})|\boldsymbol{e},\mathcal{C}(NOI))} = \\ \frac{p(I=\mathcal{C}(\mathcal{I}))}{1-p(I=\mathcal{C}(\mathcal{I}))} \times \\ \prod_{j=1}^{N} \frac{p(\boldsymbol{e}_j|I=\mathcal{C}(\mathcal{I}),\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j-1})}{p(\boldsymbol{e}_j|I\neq\mathcal{C}(\mathcal{I}),\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j-1})}\end{gathered} \tag{4.13}$$

The last factor in Equation 4.13 is the product of the conditional likelihood ratios, where the conditional likelihood ratio is given by the following equation:

$$\begin{gathered}LR\left(\mathcal{C}\left(\mathcal{I}\right) : \boldsymbol{e}_j|\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j-1}\right) = \\ = \frac{p\left(\boldsymbol{e}_j \mid I=\mathcal{C}\left(\mathcal{I}\right),\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j-1}\right)}{p\left(\boldsymbol{e}_j \mid I\neq\mathcal{C}\left(\mathcal{I}\right),\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j-1}\right)}.\end{gathered} \tag{4.14}$$

The conditional likelihood ratio $LR\left(\mathcal{C}\left(\mathcal{I}\right) : \boldsymbol{e}_j|\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j-1}\right)$ quantifies the contribution of the j^{th} agent's findings $\boldsymbol{e}_j$ in the population evidence $\boldsymbol{e} = \{\boldsymbol{e}_1, \boldsymbol{e}_{2,} \ldots, \boldsymbol{e}_N\}$ to the probability of instantiation $\mathcal{I} = \mathcal{C}\left(\mathcal{I}\right)$ of interface nodes given the background information represented by evidence set $\mathcal{B}_{j-1} = \boldsymbol{e}_1, \ldots, \boldsymbol{e}_{j-1}$. Using $\beta_{j-1}$to denote background information, the likelihood ratio is given by following equation:

$$\begin{gathered}LR\left(\mathcal{C}\left(\mathcal{I}\right) : \boldsymbol{e}_j|\mathcal{B}_{j-1}\right) = \\ = \frac{p\left(\boldsymbol{e}_j \mid I=\mathcal{C}\left(\mathcal{I}\right),\mathcal{B}_{j-1}\right)}{p\left(\boldsymbol{e}_j \mid I\neq\mathcal{C}\left(\mathcal{I}\right),\mathcal{B}_{j-1}\right)}.\end{gathered} \tag{4.15}$$

The score that I plan to use to select agents with evidence that supports an instantiation of interface nodes is given by Equation 4.16.

$$score\left(\mathcal{C}\left(\mathcal{I}\right) : \boldsymbol{e}_j|\mathcal{B}_{j-1}\right) = LR\left(\mathcal{C}\left(\mathcal{I}\right) : \boldsymbol{e}_j|\mathcal{B}_{j-1}\right). \tag{4.16}$$

As can now be seen from Equation 4.14, the contribution of each agent depends on the evidence in $\mathcal{B}$ that we are conditioning on and therefore is based on the order in which we analyze evidence. I will use a best-first approach to determine the order in which the evidence will be analyzed as j^{th} in the order, given the previous $j-1$ agents are selected as represented by set $\mathcal{B}_{j-1} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{j-1}\}$. The j^{th} selected evidence $\boldsymbol{e}_{s_j}$ is chosen to satisfy the following condition:

$$\boldsymbol{e}^*_{s_j} = \max_{\boldsymbol{e}_{s_j} \in \{\boldsymbol{e}_j, \ldots, \boldsymbol{e}_N\}} \arg \left(LR\left(\mathcal{I} = \mathcal{C}\left(\mathcal{I}\right) : \boldsymbol{e}_{s_j} | \mathcal{B}_{j-1}\right)\right). \tag{4.17}$$

Once the j^{th} evidence $\boldsymbol{e}^*_{s_j}$ is selected, it is added to the background evidence set $\mathcal{B}_{j-1}$, thus creating set $\mathcal{B}_j$. In this way we are first selecting high-ranked evidence, and each piece of evidence is scored conditionally on the previously selected high-ranked evidence. I denote the score calculated using the best-first approach as S^B. This way instantiations in the selected set are selected using the scores calculated conditionally on the previously selected instantiations with higher scores. If the order in which the evidence was selected does not correspond to the order given by the score, this indicates that scores of such evidence are sensitive to the background evidence set $\mathcal{B}$.

A symmetric alternative to best-first scoring is a worst-first approach, which always first adds the evidence with the lowest score to the set of scored evidence. Next, scores for the remaining evidence are recalculated conditionally on the set of scored evidence and the evidence with the lowest score is added to the set of scored evidence. This continues until all evidence is in the set of scored evidence. The score of the last added piece of evidence $\boldsymbol{e}_{last}$ is calculated conditionally on all the remaining evidence $\boldsymbol{e} \setminus \boldsymbol{e}_{last}$. Therefore the j^{th} selected evidence $\boldsymbol{e}_{s_j}$ satisfies the following condition:

$$\boldsymbol{e}_{s_j} = \arg\min_{\boldsymbol{e}_{s_j}} \left(LR\left(\mathcal{I} = i : \boldsymbol{e}_{s_j} | \mathcal{B}_{j-1}\right)\right), / \tag{4.18}$$

where $i \in \mathcal{C}\left(\mathcal{I}\right)$. I denote the score calculated using the worst-first approach as S^W. I can identify some pieces of evidence $\boldsymbol{e}$ that are sensitive to the order in which the score is calculated by comparing the scores obtained using the best-first and the worst-first methods. This heuristic approach aims to identify the pieces of evidence whose scores are sensitive to the evidence set $\mathcal{B}$. Such pieces of evidence can be marked as sensitive in the explanation.

These two sets of scores (obtained using the best-first and the worst-first methods) could be complemented with scores calculated using different, randomly chosen, alignments of evidence. This is, however, a computationally more expensive solution.

All of the scoring strategies described above satisfy decomposition of the total score, which is given by the equation

$$S\left(i:e\right)=\prod_{j=1}^{N}S\left(i:e_j|\mathcal{B}_{j-1}\right). \tag{4.19}$$

This property is used for a graphical analysis of the evidential support for an instantiation i. I refer to this property as the decomposition property.

Another way to score the evidence of agents is based on an analogy to INSITE's cost of commission and cost of omission measures (Section 3.3.4). However, the individual scores calculated using this method cannot be combined to obtain a score for the total available evidence, as shown by Equation 4.19. In Section 4.2.6 I discuss how the various evidence scoring methods mentioned above can be used to select the evidence for the explanation.

Although multivalued interface nodes make explanation more difficult, the LR can provide valuable information. Since WOE is a monotonic transformation of LR, LR allows us to determine which evidence supports and which evidence contradicts the instantiation $\mathcal{C}\left(\mathcal{I}\right)$ in the same way WOE does (Equation 4.7). This is demonstrated in Equation 4.20.

$$LR(\boldsymbol{i}:\boldsymbol{e_j}|\mathcal{B}_{j-1})=\begin{cases}>1 & \text{evidence } e_j \text{ supports instantiation } i\\ & \text{over instantiation } \neg i \text{ in the context of } \mathcal{B}\\ =1 & \text{evidence } e_j \text{ supports instantiation } i\\ & \text{as much as instantiation } \neg i \text{ in the context of } \mathcal{B}\\ <1 & \text{evidence } e_j \text{ supports instantiation } \neg i\\ & \text{over instantiation } i \text{ in the context of } \mathcal{B}.\end{cases}. \tag{4.20}$$

Moreover, LR allows us to say that the instantiation i is $LR(\boldsymbol{i}:\boldsymbol{e_j}|\mathcal{B}_{j-1})$ times more (or alternatively, less) supported by e_j than $\neg i$ given the background information $\mathcal{B}_{j-1}$ if $LR(\boldsymbol{i}:\boldsymbol{e_j}|\mathcal{B}_{j-1})>0$ (or alternatively, $LR(\boldsymbol{i}:\boldsymbol{e_j}|\mathcal{B}_{j-1})<0$).

In the next section I describe the criteria I plan to use to specify how many instantiations of *NOI*, *I*, and the agents' evidence will be included in an explanation.

4.2.6 Selecting and clustering information for explanation

Simplification of the explanation In general, improvement and simplification of an explanation can be achieved by presenting only relevant information. Instead of including all instantiations of the nodes of interest (NOI), interface nodes ($\mathcal{I}$), and evidence nodes, only instantiations that most support and contradict nodes on the immediate higher level will be included in a hierarchical explanation. The hierarchical explanation is constructed from the *NOI* (top level) through the interface nodes (middle level) to the agent networks and agent evidence (bottom level). The importance of each instantiation on the given level, conditional on the configuration of the higher level, is measured using the score given by Equation 4.1 for *NOI*, by Equation 4.2 for interface nodes, and by Equation 4.16 for the evidence observed for agents in the population.

Selecting instantiations and evidence In general, there may be many configurations at each level. We want to include only the most useful information in the explanation.

First, we start by selecting instantiations of the NOI. Once the scores of the instantiations are calculated using Equation 4.1, the instantiations are sorted in descending order, forming the ordered set $\mathcal{C}^D(NOI)$. The superscript D in $\mathcal{C}^D(NOI)$ denotes descending order of instantiations in $\mathcal{C}^D(NOI)$ with respect to their score. In order to reduce the number of instantiations in the explanation, HEM selects subset $\mathcal{C}^D_{\theta_{NOI}}(NOI)$ using threshold θ_{NOI}, which represents the minimal relative share of the cumulative score for the selected instantiations on the cumulative score for all instantiations of NOI. Since the cumulative score for all instantiations of NOI is 1, this condition can be expressed as:

$$min\left(\left|\mathcal{C}^D_{\theta^+_{NOI}}(NOI)\right|\right) \text{ such that } \sum_{noi \in \mathcal{C}^D_{\theta^+_{NOI}}(NOI)} S(NOI = \text{noi}) \geq \theta_{NOI} \tag{4.21}$$

and instantiations in $C^D_{\theta^+}$are in the same order as in the ordered set $C^D\,(NOI)$

where $\left|\mathcal{C}^D_{\theta_{NOI}}\,(NOI)\right|$ denotes the number of entries in the set $\mathcal{C}^D_{\theta_{NOI}}\,(NOI)$. If the last evidence added to a set $C^D_{\theta^+}$ has the same score as the next evidence in C^D, then the set $C^D_{\theta^+}$ specified by the above criteria is not unique. For the purpose of explanation I will consider these sets to have equivalent explanatory power and any of them can be presented as an explanation. Another alternative is to require that every configuration in the selected set $\mathcal{C}^D_{\theta_{NOI}}\,(NOI)$ has at least a minimal score S, given by threshold θ^+_{NOI} :

$$\forall noi \in \mathcal{C}^D_{\theta_{NOI}}\; S(NOI = \text{noi}|\boldsymbol{e}) > \theta_{NOI}\,. \tag{4.22}$$

The user can control how much information is included in an explanation by choosing a threshold θ, which regulates how many instantiations are included in the set $\mathcal{C}^D_\theta$. It is also possible to limit the number of configurations in $\mathcal{C}^D_\theta$ with the chosen count c_{NOI} (Equation 4.23).

$$\left|\mathcal{C}^D_{\theta_{NOI}}\,(NOI)\right| \leq c_{NOI}\,. \tag{4.23}$$

All criteria described by Equations 4.23, 4.22, and 4.21 can be combined to create new criteria for selection of instantiations.

Next, HEM proceeds to selection of the states of the joint interface nodes ($\mathcal{I}$) for each configuration selected in the previous step. The same approach which was used for configurations of the NOI can be applied to selection of configuration of interface nodes, with the scores given by Equation 4.2 and threshold $\theta_\mathcal{I}$, where $noi \in \mathcal{C}^D_{\theta_{NOI}}\,(NOI)$. The selection process results in set $\mathcal{C}^D_{\theta_\mathcal{I}}\,(\mathcal{I})$ of the most important configurations of interface nodes, given the configuration $NOI = noi$.

Finally, HEM selects the agents in the population with evidence $\boldsymbol{e}_j$ that contributes most to the posterior probability of configuration $i \in \mathcal{C}^D_{\theta_\mathcal{I}}\,(\mathcal{I})$. To do this, it uses the score from Equation 4.16, clustering and restricting the number of patient groups, represented by equivalence classes, which will be included in the explanation. Since the score for selection of agents with evidence supporting state $i \in \mathcal{C}^D_{\theta_\mathcal{I}}\,(\mathcal{I})$ of the interface node is based on a likelihood ratio, we can use the evidence to split agents in the population into 3 basic

groups (from Equation (4.20)): agents with *supporting* ($S(i:\boldsymbol{e}_j|\mathcal{B}_{j-1}) \gg 1$), *contradicting* ($S(i:\boldsymbol{e}_j|\mathcal{B}_{j-1}) \ll 1$) and *neutral* ($S(i:\boldsymbol{e}_j|\mathcal{B}_{j-1}) \approx 1$) evidence with regard to configuration i. Instantiations of $\mathcal{I}$ that have the score ($S(i:\boldsymbol{e}_j|\mathcal{B}_{j-1}) \approx 1$) do not individually either substantially support or contradict hypothesis $\mathcal{I} = i$. In order to simplify the explanation, HEM aggregates these neutral findings and represents them with the label "neutral evidence" and an aggregated score. In this way we want to reduce the amount of neutral evidence in the explanation (score $S(\mathcal{I} = i \mid NOI = S(NOI|\mathbf{e}), \mathbf{e}) \approx 1$). Parameters $\theta^+_{\mathbf{E}}$ and $\theta^-_{\mathbf{E}}$ are used to control how many neutral findings will be filtered out. $\theta^+_{\mathbf{E}}$ and $\theta^-_{\mathbf{E}}$ represent the minimum and the maximum score of findings in the group of supporting or contradicting evidence. Findings $\boldsymbol{e}_j$ that satisfy the condition $S(i:\boldsymbol{e}_j|\mathcal{B}_{j-1}) \geqq \theta^+_{\mathbf{E}}$ for score S are selected for the explanation as evidence supporting the configuration $i \in \mathcal{C}(\mathcal{I})$,and findings $\boldsymbol{e}_j$ that satisfy condition $S(i:\boldsymbol{e}_j|\mathcal{B}_{j-1}) \leqq \theta^-_{\mathbf{E}}$ are selected for the explanation as contradicting it. Findings e_j that satisfy $\theta^+_{\mathbf{E}} > S(i:\boldsymbol{e}_j|\mathcal{B}_{j-1}) > \theta^-_{\mathbf{E}}$ will be labeled as neutral and presented in condensed form in the explanation. We can also add criteria pertaining to the maximum number of selected supporting and contradicting findings to the constraints for selecting neutral evidence. The same approach can be applied to evidence consisting of a group of findings rather than from a single finding (Section 4.2.6).

Clustering of instantiations and evidence Clustering states of variables and similar agents in the population and explaining them together is a useful way to simplify the explanation.

In order to cluster instantiations of the NOI or instantiations of interface nodes, first the ordered sets $\mathcal{C}^D(NOI)$ and $\mathcal{C}^D(\mathcal{I})$ are created. In the case where NOI or $\mathcal{I}$ have many possible instantiations with similar scores, similar instantiations in $\mathcal{C}^D$ could be clustered together and explained in the next step as a group. I consider instantiations to be similar if they have a similar importance for the explanation. Consider the clustering of instantiations of interface nodes. The simplest way to partition the sets of selected instantiations is to use predefined thresholds $\theta_1, \ldots, \theta_R$ to partition the set of instantiations C^D into ordered

subsets C^{D1}, $C^{D2}, \ldots, C^{DR}$ using the corresponding cumulative score $S^r(\mathcal{I})$ for $C^{Dr}(\mathcal{I})$:

$$S^r(\mathcal{I}) = \sum_{i \in C^{Dr}} S(\mathcal{I} = i)\,. \tag{4.24}$$

Using cumulative scores $S^1(\mathcal{I}), \ldots, S^r(\mathcal{I}), \ldots, S^R(\mathcal{I})$, C^D can be partitioned into C^{D1}, $C^{D2}, \ldots, C^{DR}$ in the following manner:

$$1 \geq S^1(\mathcal{I} = i) \geq \theta_1,\ \theta_1 \geq S^2(\mathcal{I} = i) \geq \theta_2, \ldots,\ \theta_{R-1} \geq S^R(\mathcal{I} = i) \geq \theta_R \qquad \text{such that} \tag{4.25}$$

$$\forall i_{r-1} \in \mathcal{C}^{D(r-1)},\ \forall k \in \mathcal{C}^{Dr}(\mathcal{I})\ \ S(\mathcal{I} = i_{r-1}) \geq S(\mathcal{I} = i_r) \tag{4.26}$$

where each $\theta \in [0, 1]$. The second condition means that the resulting subsets C^{D1}, $C^{D2}, \ldots, C^{DR}$ represent an ordered partition of $\mathcal{C}^D$. An ordered partition of $\mathcal{C}^D(NOI)$ can be obtained in the same way. This, however, does not necessarily group instantiations with similar scores. Therefore a better way is to use clustering algorithms such as k-means to partition the scores. Clustering algorithms, however, usually require an input parameter that specifies the number of clusters. The number of clusters can be estimated using available methods (Sugar and James, 2003; Tibshirani et al., 2001).

For clustering agents in the population, I use equivalence class Q_k^l to represent agents that have an identical pair $\langle M_l, e_k \rangle$, where e_k is evidence observed for the agent and M_l is the agent's model. The symbol N_k^l will be used to designate the count of the agents in the equivalence class Q_k^l. Let Ω be the set of all equivalence classes such that total evidence can be partitioned based on equivalence classes Q_k^l so that:

$$\boldsymbol{e} = \bigcup_{\langle M_l, \boldsymbol{e}_k \rangle \in \Omega} (\boldsymbol{e}_k)^{N_k^l}\,, \tag{4.27}$$

where exponent N_k^l symbolizes that we are including N_k^l copies of evidence $\boldsymbol{e}_k$ for class Q_k^l. The total observed evidence can be partitioned using the equivalence classes as shown below.

$$P(\boldsymbol{e}|\mathcal{I} = i) = \prod_{Q_k^l \in \Omega} P(E = \boldsymbol{e}_k|\mathcal{I} = i)^{N_k^l}\,, \tag{4.28}$$

where $Q_k^l = \langle M_l, e_k \rangle$. However, Equation 4.28 is correct only if configuration i consists of only one state of $\mathcal{I}$. This is due to the loss of independence of agents which results if we condition on more than one configuration of $\mathcal{I}$. Let $e_j^{Q_k^l}$ represent the evidence of the j^{th}

agent in the evidence equivalence class Q_k^l. If i consists of more than two instantiations, the posterior probability of evidence $\boldsymbol{e}$ must be partitioned according to Equation 4.29.

$$P\left(\boldsymbol{e}|\mathcal{I}=i\right) = \prod_{\substack{Q_k^l\in\Omega \\ 0\leq k\leq K \\ 0<l<L}} \prod_{\substack{e_j^{Q_k^l}\in Q_k^l \\ j=1,\ldots,N_k^l}} P\left(E=\boldsymbol{e}_j^{Q_k^l}|\mathcal{I}=i,\boldsymbol{e}_1^{Q_1^1},\ldots,\boldsymbol{e}_{N_1^1}^{Q_1^1},\ldots,\boldsymbol{e}_1^{Q_k^l},\ldots,\boldsymbol{e}_{j-1}^{Q_k^l}\right) \quad (4.29)$$

The simplest case occurs if all agents in the population have identical networks, i.e., $M_j = M$ for all $j = 1,\ldots,N$, and the same variables are observed as evidence for every agent in the population, i.e., $E_j = E$ for $j = 1,\ldots,N$. In such cases it is possible to ignore the upper index and write the equivalence class as $Q_k = \left\{\forall\text{agents } j, \text{ such that } \langle M_j,\boldsymbol{e_j}\rangle = \left\langle M,\boldsymbol{e}^{Q_k}\right\rangle\right\}$. These equivalence classes, then, differ only in evidence. I will refer to them as *evidence equivalence classes*. Let N_k equal the number of agents with evidence $\boldsymbol{e}^{Q_k}$. These agents constitute evidence equivalence class Q_k . If the ratio of the patient count and number of possible instantiations of evidence nodes is high, it is efficient to group agents in the population into equivalence classes by the observed evidence. A given piece of evidence e for an entire population corresponds to a unique set Ω of equivalence classes. With equivalence classes, the quantity $P\left(\boldsymbol{e}|\mathcal{I}=i\right)$ is given by the expression

$$P\left(\boldsymbol{e}|\mathcal{I}=i\right) = \prod_{Q_j\in\Omega} P\left(E=\boldsymbol{e}_{Q_k}|\mathcal{I}=i\right)^{N_k}, \quad (4.30)$$

where N_k is the instance count of equivalent class Q_k. Let $e_j^{Q_k}$ represent the evidence of the j^{th} agent in the evidence equivalence class Q_k. Similar to Equation 4.28, Equation 4.30 is valid only if configuration i consists of only one state of $\mathcal{I}$. If i consists of more than one instantiation of $\mathcal{I}$, $P\left(\boldsymbol{e}|\mathcal{I}=i\right)$ must be calculated using the following equation:

$$P\left(\boldsymbol{e}|\mathcal{I}=i\right) = \prod_{\substack{Q_k\in\Omega \\ 0\leq k\leq K}} \prod_{\substack{e_j^{Q_k}\in Q_k \\ j=1,\ldots,N_k}} P\left(E_j=\boldsymbol{e}_j^{Q_k}|\mathcal{I}=i,\boldsymbol{e}_1^{Q_1},\ldots,\boldsymbol{e}_{N_1}^{Q_1},\boldsymbol{e}_1^{Q_2},\ldots,\boldsymbol{e}_1^{Q_k},\ldots,\boldsymbol{e}_{j-1}^{Q_k}\right) \quad (4.31)$$

The score for the evidence provided by an equivalence class can be calculated similarly to the score for the agent's evidence. Explanation of interface nodes using the evidence of a whole equivalence class instead of the evidence of individual agents simplifies the explanation and, under certain conditions, enables faster calculation of the explanation. Let $\boldsymbol{e}_{Q_k} =$

$\left\{\boldsymbol{e}_1^{Q_k}, \ldots, \boldsymbol{e}_{N_k}^{Q_k}\right\}$be the total evidence of the evidence equivalence class Q_k, where $\boldsymbol{e}_j^{Q_k} = \boldsymbol{e}^{Q_k}$ for $j = 1, \ldots, N_k$. Equation 4.14 can be rewritten in terms of equivalence classes as shown in Equation 4.32.

$$\begin{aligned} eq : LR_Dependent_E QLR\left(\mathcal{C}\left(\mathcal{I}\right) : \boldsymbol{e}_{Q_k} | \boldsymbol{e}_{Q_1}, \ldots, \boldsymbol{e}_{Q_{k-1}}\right) = \\ = \frac{p\left(\boldsymbol{e}_{Q_k} \mid \mathcal{I} = \mathcal{C}\left(\mathcal{I}\right), \boldsymbol{e}_{Q_1}, \ldots, \boldsymbol{e}_{Q_{k-1}}\right)}{p\left(\boldsymbol{e}_{Q_k} \mid \mathcal{I} \neq \mathcal{C}\left(\mathcal{I}\right), \boldsymbol{e}_{Q_1}, \ldots, \boldsymbol{e}_{Q_{j-1}}\right)} . \end{aligned} \tag{4.32}$$

If neither $C\left(\mathcal{I}\right)$, nor $\neg C\left(\mathcal{I}\right)$ represents a single instantiation, we have to calculate the numerator and denominator of Equation 4.32 using Equations 4.33 and 4.34.

$$p\left(e_{Q_j} | \mathcal{I} = \mathcal{C}\left(\mathcal{I}\right), \boldsymbol{e}_{Q_1}, \ldots, \boldsymbol{e}_{Q_{j-1}}\right) = \prod_{j=1}^{N_k} p\left(e_j^{Q_k} | \mathcal{I} = \mathcal{C}\left(\mathcal{I}\right), \boldsymbol{e}_{Q_1}, \ldots, \boldsymbol{e}_{Q_{j-1}}, \boldsymbol{e}_1^{Q_k}, \ldots, \boldsymbol{e}_{j-1}^{Q_k}\right) . \tag{4.33}$$

$$p\left(e_{Q_j} | \mathcal{I} \neq \mathcal{C}\left(\mathcal{I}\right), \boldsymbol{e}_{Q_1}, \ldots, \boldsymbol{e}_{Q_{j-1}}\right) = \prod_{j=1}^{N_k} p\left(e_j^{Q_k} | \mathcal{I} \neq \mathcal{C}\left(\mathcal{I}\right), \boldsymbol{e}_{Q_1}, \ldots, \boldsymbol{e}_{Q_{j-1}}, \boldsymbol{e}_1^{Q_k}, \ldots, \boldsymbol{e}_{j-1}^{Q_k}\right) . \tag{4.34}$$

4.2.7 Various treatments of the interface node

There are two interface nodes in PANDA-CDCA: *Outbreak Disease in Population* and *Fraction of Population with Outbreak Disease and ED Visit.* I will refer to these nodes as *Outbreak Disease* (D_O) and *Fraction of Population* (F). The treatment of interface nodes described in Sections 4.2.4 and 4.2.6 assumes equal importance of nodes for the explanation. However, interface nodes are not equally important for explanation of inference. Assume that *Outbreak Disease* provides more valuable information about the outbreak for the user than *Fraction of population.* There is an alternative way to treat nodes with a different potential contribution to the explanation. One way is to completely disregard frequency and consider $\mathcal{I} = \{Outbreak\, Disease\}$ in the equation for calculating scores and selecting and explaining instantiations (Equation 4.35).

$$S\left(D_O\right) = P\left(D_O | O, \boldsymbol{e}\right) . \tag{4.35}$$

Another alternative includes all interface nodes in the explanation, but starts with the most informative interface node. First, the explanation for $\mathcal{I} = \{Outbreak\,Disease\}$ is generated as in the previous approach (Equation 4.35). In the next step, instantiations of *Fraction of Population* are explained conditionally on the instantiations D_O selected in the previous step. The score of instantiation f of node F conditional on the state d_O of the node D_O can be calculated using Equation 4.36.

$$S(F) = P(F|D_O, O, \boldsymbol{e}). \tag{4.36}$$

4.2.8 User interface for presenting a hierarchical explanation

The explanation generated (Section 4.2.3) will be presented in the simple *user interface* shown in Figure 22 on page 72. By default, HEM displays the explanation consisting of the instantiations and evidence with the highest explanatory power. The user interface will allow the user to select other instantiations and agent evidence with a left click of the mouse.

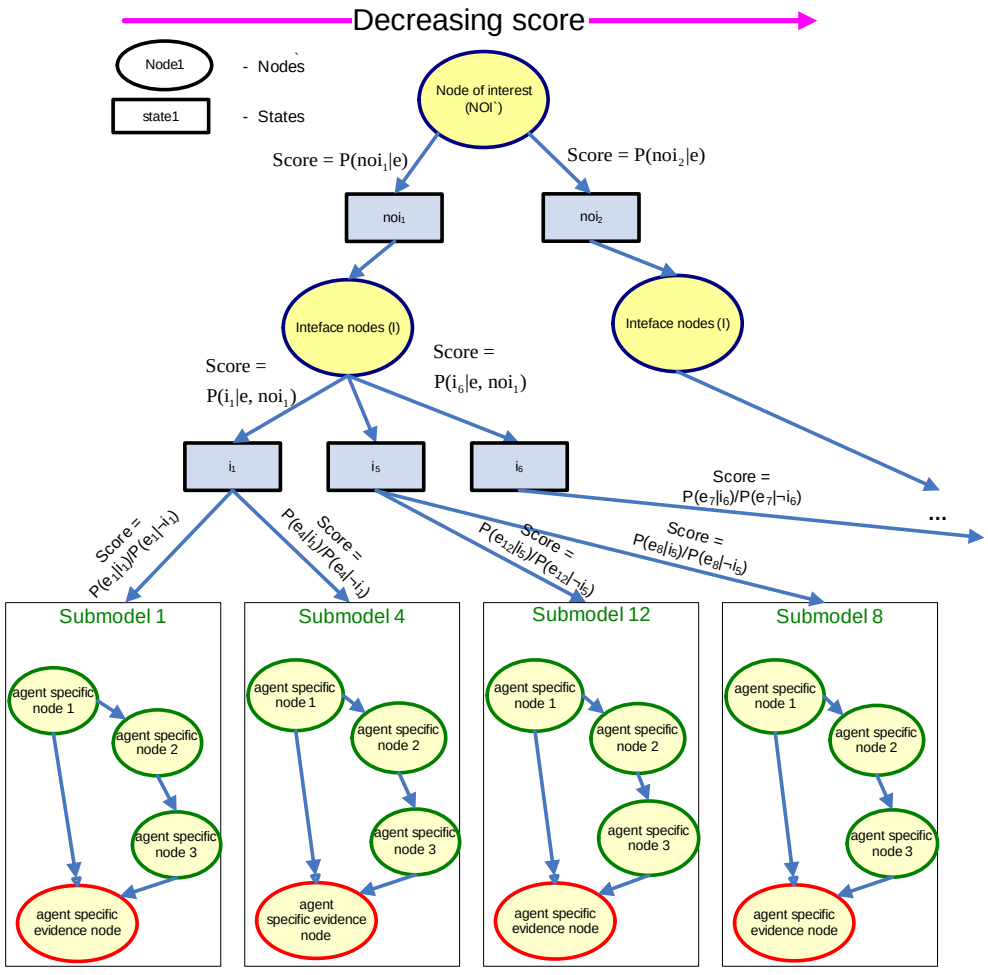


Figure 18: General schema of hierarchical explanation

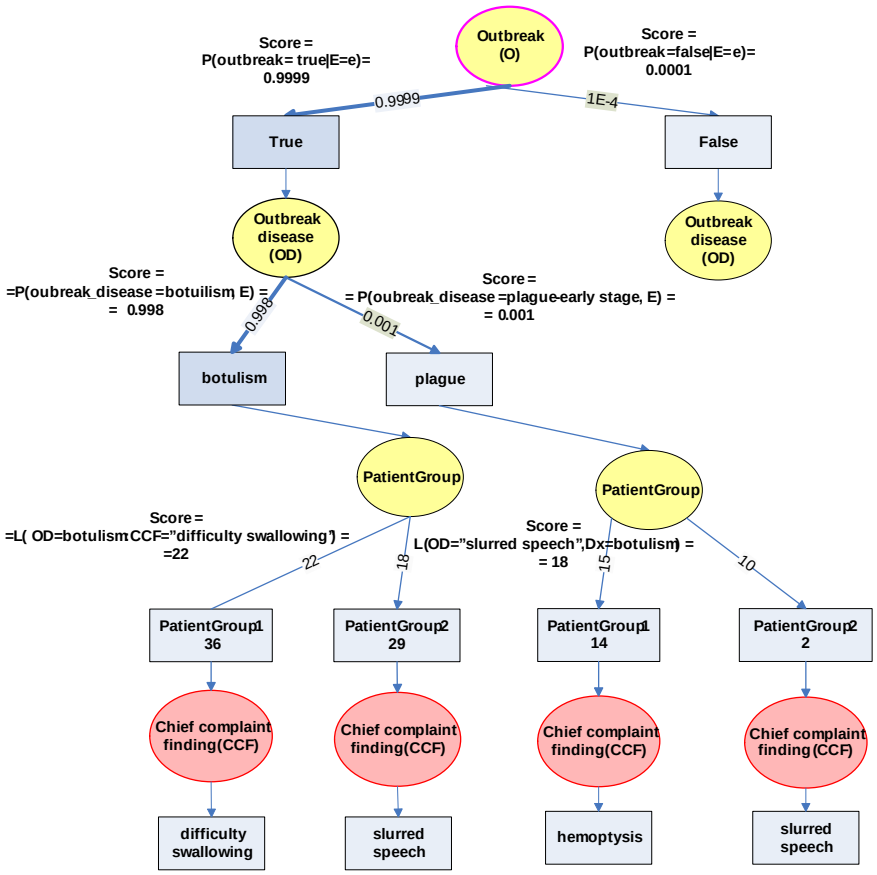


Figure 19: Schema of hierarchical explanation for PANDA-CDCA. Numbers shown for patient groups represent counts of patients belonging to the patient group.

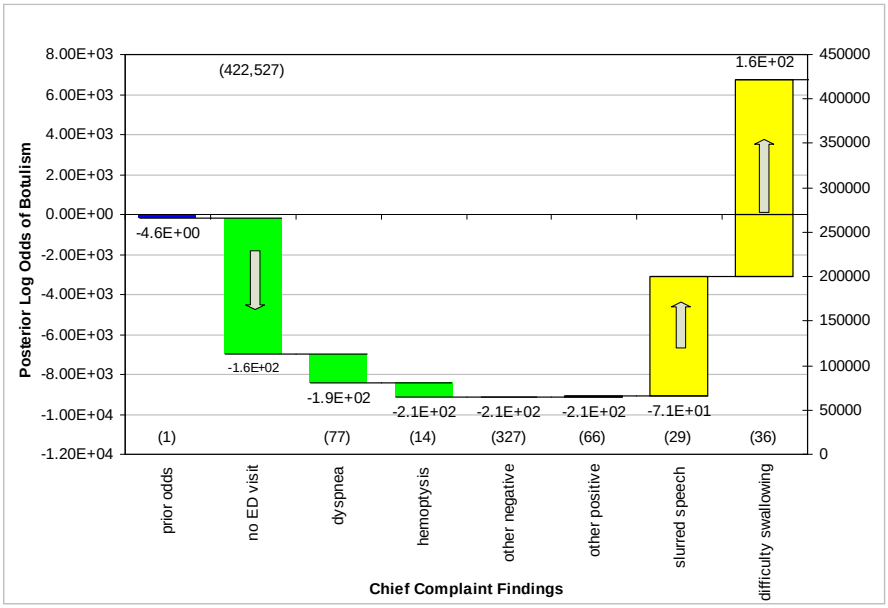


Figure 20: Graphical analysis of evidence. Impact of evidence on posterior odds of botulism. Numbers alongside each bar represent the final odds due to priors, current evidence, and evidence to the left of each bar. Numbers in parentheses represent the number of agents in the evidence equivalence group.

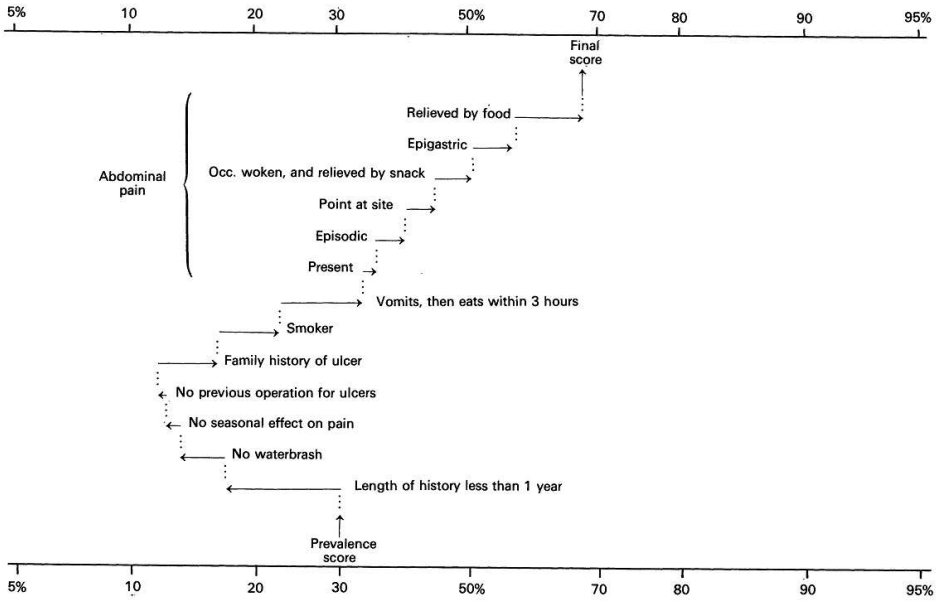


Figure 21: Fully graphical explanation that is using weight of evidence from balance sheet in Figure 9 to explain contribution of the findings in evidence to posterior probability of peptic ulcer (Spiegelhalter and Knill-Jones, 1984).

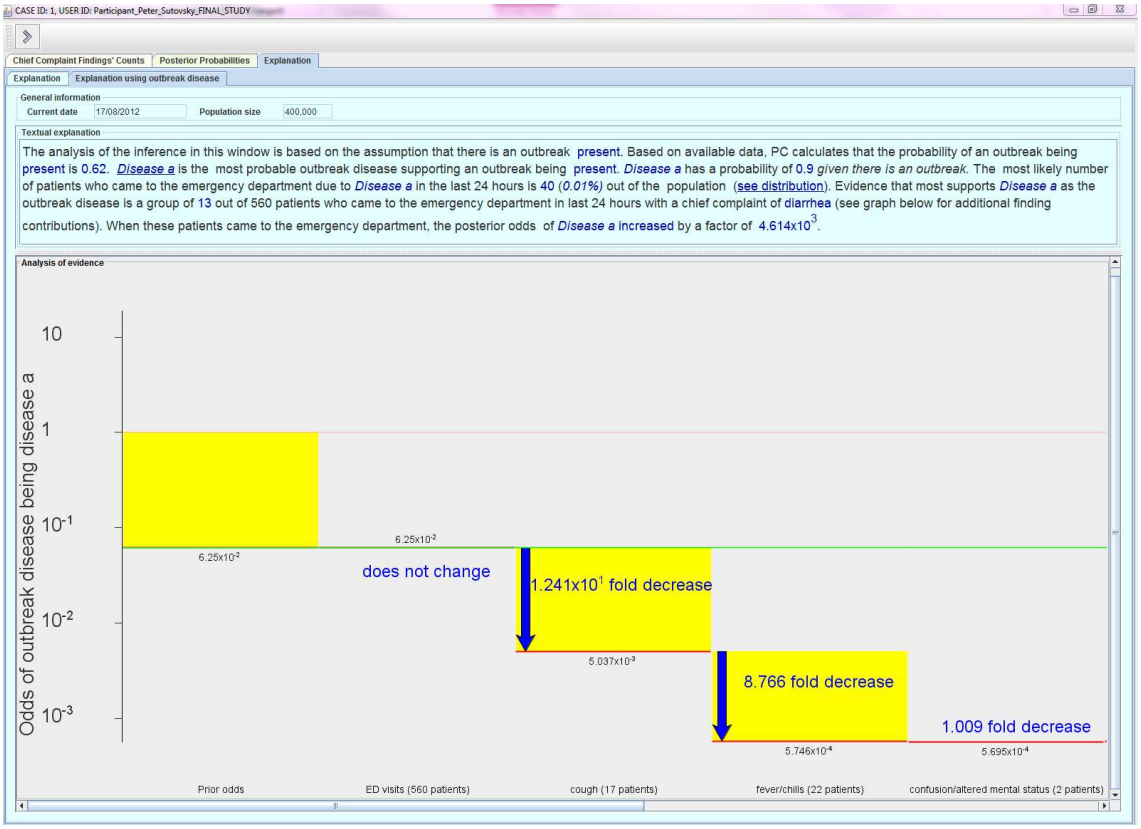


Figure 22: User interface for HEM.

5.0 EXPERIMENTAL EVALUATION

5.1 INTRODUCTION

In this section, I describe how I evaluated the potential utility of the proposed explanation methods. The experimental design I used is an adaptation of the experimental design used by Suermondt (1992) for evaluation of explanation methods in the medical domain.

The proposed explanation method is applicable to any Bayesian network (BN) with a population of independent agents. In order to evaluate the method, a specific BN is needed. Due to pragmatic limitations, I narrowed the scope of the study to evaluation of the effect of the HEM (Hierarchical Explanation Method) in the biosurveillance field. I chose to evaluate the explanation method using the biosurveillance domain because the lab in which I am conducting the research has developed an agent-based Bayesian network approach to surveillance. Moreover, biosurveillance is currently a very important task, due to globalization and the elevated threat of pandemic disease outbreaks. I chose to use the a simple biosurveillance network (SBN), which represents a simple Bayesian model with a population of independent agents in which submodels of agents are modeled as conditionally independent of each other given a small set of interface nodes (Section 5.2). This model enables detection of disease outbreaks caused by non-contagious diseases (see Section 5.2). SBN was created as a variation of PANDA-CDCA (Section 4.1.2) in order to simplify the domain for the experimental study participants.

In this experimental study, I investigate the objective and subjective effect of explanation on the quality of decisions made by users of this biosurveillance system based on an agent-based Bayesian network. The objective part of the evaluation involves measuring the accuracy of assessment of the health status of the population (probability of outbreak being

present, type of outbreak disease, number of infected people arriving at the hospital) and the consistency of the model with the domain knowledge provided. The subjective part of the evaluation assesses the study participant's confidence in his or her assessments.

5.2 SIMPLE BIOSURVEILLANCE NETWORK (SBN)

The *simple biosurveillance network (SBN)* illustrated in Figure 23 is a BN for a scenario in which the chief complaint findings of patients coming to emergency departments (EDs) in hospitals in a monitored area are recorded and reported to a biosurveillance system. The structure of the SBN model is identical to that of PANDA-CDCA. In order to simplify the domain, however, SBN represents only a few findings (confusion/altered mental status, cough, diarrhea, fever/chills, other findings, and other outbreak findings) and three outbreak diseases (cryptosporidiosis, influenza, and hepatitis A). "Other findings" represents disjunctively findings that were not important for any of the diseases modeled in PANDA-CDCA and therefore were included in the model as a group representing any of one them. When I reduced the number of findings to create the SBN, I created another group, "other outbreak findings", to represent findings that occur with outbreak diseases but were not modeled as individual findings for the sake of simplicity. Simplification of the domain to make it more comprehensible to users was motivated by the practical need to test HEM with study participants lacking substantial knowledge of biosurveillance. I use generic names for diseases (e.g., disease-A and disease-B) instead of real names (Table 3) to avoid (1) user intimidation by an unknown domain and (2) possible conflict of a participant's current knowledge of a disease with the SBN's model of the disease. For this study I consider SBN to be a correct model of the domain.

Since I want to evaluate whether explanation can help users make the correct decision even in cases where the system provides incorrect advice, I need a second model that provides incorrect advice for some scenarios. I created a biased model as a variation of the SBN, which I refer to as SBN-B (SBN-Biased). SBN-B was obtained from the SBN by changing the

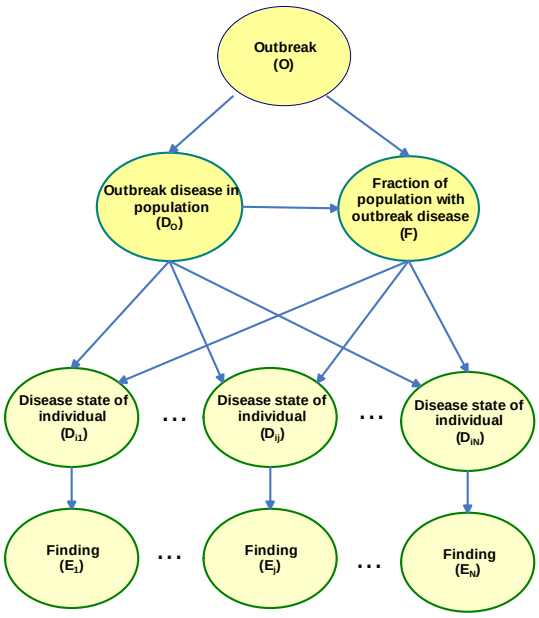


Figure 23: Simple biosurveillance network

conditional probabilities of the chief complaint findings for "hepatitis A" such that SBN-B provides incorrect advice for scenarios with an outbreak of "hepatitis A".

5.3 METHODS

In order to evaluate the explanation method, I used simulated information about population health that consisted of records of the findings in patients who visited the emergency department. One scenario consists of daily findings collected for the monitored segment of the population by the emergency department over a certain period of time prior to the assessment time. Study participants were asked to assess the outbreak-disease status of the population at the end of this period. The simulated records used in the evaluation contain the time and the chief complaint finding. I evaluated HEM with 18 graduate students, 3 undergraduate students and one faculty member, as they are easier to enlist in a study than the experienced public health officials who are the potential end users of a biosurveillance

Table 3: Example of states of the nodes in the SBN

Node	States
Outbreak (O)	true, false
Outbreak disease in population (D_o)	disease_A, disease_B, no disease
Fraction of population with outbreak disease (F)	$10^{-3}, 10^{-4}, \ldots$
Disease state of the j^{th} individual (D_{ij})	disease_A, disease_B, no disease
Findings of j^{th} individual (E_j)	finding_A, finding_B, finding_A and finding_B, ...

system. Moreover, since the public health professionals reside in various locations throughout the country, their participation would make the evaluation much more complicated and time-consuming. The participants were expected to have sufficient knowledge of probability. In addition, the complexity of the domain was adjusted to match the limited experience of the participants in using a large amount of evidence to make inferences.

Study participants were presented with several outbreak and non-outbreak scenarios. *Simple Biosurveillance System-Correct* (SBS-C) is the system based on the SBN, and *Simple Biosurveillance System-Biased* (SBS-B) is the system based on the SBN-B (Figure 24). At the top of SBS-C and SBS-B is the *Simple Biosurveillance System* (SBS), which provides study participants with the correct predictions using SBS-C or the incorrect predictions using SBS-B (Figure 24). I compared assessments by study participants who see only SBS conclusions with no explanations (control mode) versus assessments by study participants who saw SBS conclusions together with explanations for the conclusions provided by HEM (intervention mode). In addition, I asked all participants to evaluate scenarios using only daily counts of patients who arrived at the ED with the selected chief complaint findings during the previous 24 hours (baseline mode).

At the beginning of the experimental evaluation, participants learned how to use the system during a short training session. During this session they were asked to evaluate one scenario in the baseline, control, and intervention modes. This also allowed participants

to familiarize themselves with the domain knowledge represented by the generative model (SBS-C). At any time during the study, the participants had access to information about the generative model, such as the prior probability of diseases and the conditional probability of seeing the findings for a given disease.

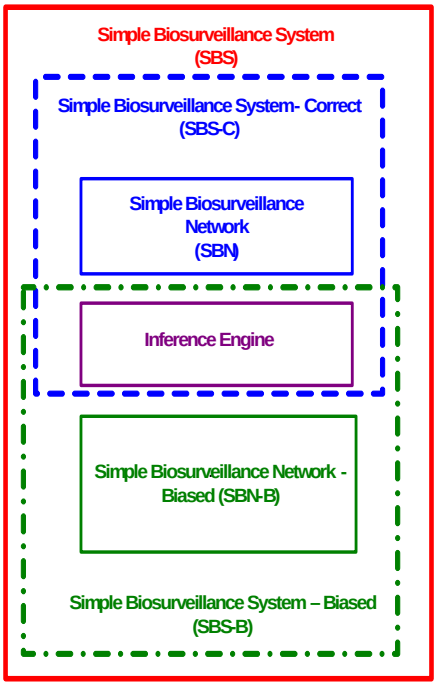


Figure 24: Schema of Simple Biosurveillance System (SBS)

The study follows a scenario-by-scenario test-retest design. First, study participants evaluated each scenario without using SBS (baseline mode), using only the daily patient counts with specific chief complaint findings over a period of about 11 days, presented as a time series. Evaluation of the scenario included an assessment of the probability that an outbreak is present, which disease is the outbreak disease, the number of patients who arrived at the ED infected with the outbreak disease in the last 24 hours, and the probability of an outbreak. Participants were provided with information about modeled diseases that allowed them to diagnose the disease modeled in SBN. After assessment of all cases in baseline mode was completed, the experiments continued with a follow-up assessment, in which each of the study participants evaluated half of the scenarios with only SBS (control mode), and the other half of the scenarios using SBS and an explanation provided by HEM

(intervention mode) (Figure 25). After the study participants had evaluated each scenario in the control or intervention mode, they were asked to characterize their experience and give their impressions of the explanation.

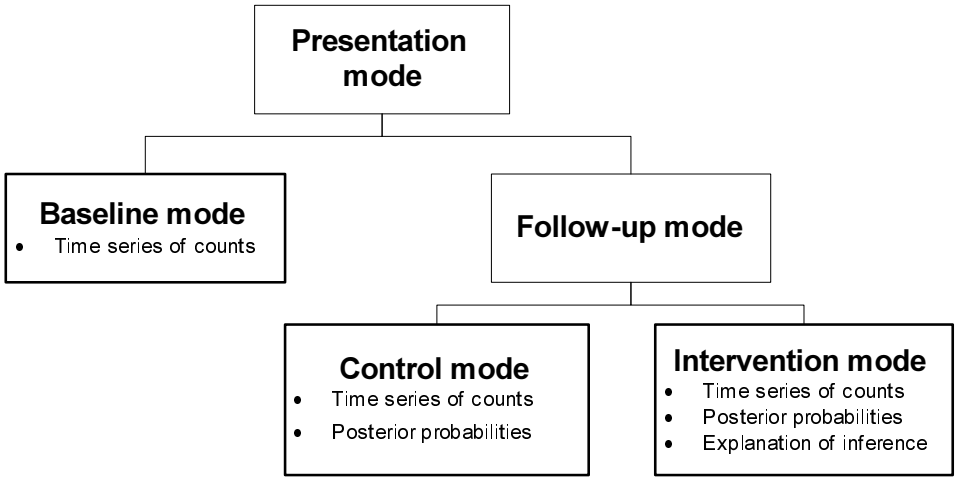


Figure 25: Computer presentation modes

The smallest component of the evaluation is the *participant scenario* (PS). A PS represents a scenario evaluated by a particular study participant. In order to measure the effect of explanation, I compared the baseline and follow-up assessments for the set of PSs assessed using only SBS with the assessments for the set of PSs assessed using SBS together with the HEM explanation facility.

5.4 GENERATION OF EVALUATION SCENARIOS

As mentioned above, I used simulated data for the experiment. I prepared a set of 16 scenarios of which 14 were outbreak scenarios and 2 were non-outbreak scenarios. Data for non-outbreak scenarios were obtained by sampling from the SBN in Figure 23, with the outbreak node O set to false. Each scenario contained daily data for a period of 11 days. In the case of an outbreak scenario, data scenarios were obtained by sampling from the SBN in Figure 23, with the *Outbreak disease* node set to a particular outbreak disease (e.g., "disease

A") and the *Fraction of Population* node set to a value corresponding to f_t. I simulated data for an outbreak scenario using epidemic models to simulate the daily fraction of patients with this outbreak disease over the time period. I always sampled N_M cases, where N_M denotes the number of people in the monitored area. I regulated the number of people that came to the ED the previous day t with the outbreak disease using parameter f_t, which denotes the fraction of such people in the whole population.

I calculated f_t for each day using the procedure described below. In order to sample scenarios for non-outbreak cases, I used a default constant fraction of the population that came to the ED with the outbreak disease ($\hat{f}$) for each day during the non-outbreak period.

I used an epidemic model to calculate the number of infected people $N_t^{I,ED}$ who came to the ED during the day t because of the outbreak disease where I in superscript denotes people infected with outbreak disease. For cryptosporidiosis and hepatitis outbreaks, I used the epidemic model developed by Eisenberg et al. (1998), and for influenza outbreaks I used a classic SIR epidemic model as developed by Kermack and McKendrick (1932).

For every day $t \in \langle 1, T \rangle$ of the outbreak, where T varied between 3 to 8 days, the patient cases were sampled from an SBN with *Outbreak disease* set to a selected outbreak disease (e.g., "disease_A") and *Fraction of Population* set to a value corresponding to ${}^{N_t^{I,ED}}/_{N_M}$. *Evidence for Fraction of Population* (F) was set such that distribution over the states $f_1, \ldots, f_{N_F}$ of F made the expected fraction of the population equal to ${}^{N_t^{I,ED}}/_{N_M}$. This means that the probability of the states $F = f_1 \ldots, F = f_{N_F}$ given by $p_{f_1}, \ldots, p_{f_N}$ must satisfy the equation

$$ {}^{N_t^{I,ED}}/_{N_M} \quad = \quad p_{f_1} f_1 + \ldots + p_{f_1} f_{N_F}. $$

Scenarios containing an outbreak begin with 3 to 8 days of non-outbreak data. The length of the non-outbreak period in the outbreak scenario was chosen depending on the length of the selected outbreak period such that the total length of a scenario was 11 days. The total length of both outbreak and non-outbreak scenarios was 11 days.

Generated scenarios were selected to fulfill the following criteria: (1) SBS-C must be able to assess the scenario with sufficient accuracy (sufficiency condition explained below), (2) the scenario must be sufficiently difficult for the participant to benefit from the decision support provided by the computer, and (3) the scenario must be simple enough for the participant

to be able to provide an assessment of an outbreak. I considered the assessment of SBS-C to be sufficiently accurate if the SBS-C using the evidence consisting of ED records on the last day of a scenario gave the highest posterior probability for the correct outbreak disease (the disease that was used to sample the data).

In practice, people are also able to look at data collected on previous days and to use the trends in the data. In order to make the experimental setting more realistic, users in the experiment were provided with historical data (11 days).

In 14 of the 16 scenarios, correct results were provided to participants as explained below. For this purpose, after a scenario was generated I tested it with the SBS-C (Section 5.3) model and verified that the computer had identified the "correct" feature (outbreak disease and fraction of population with outbreak disease) of an outbreak as the most probable one. Since the computer provided correct conclusions for scenarios, we expected increased user performance when using SBS-C in the follow-up mode as compared to the baseline mode. Improvement in user performance thus corresponded to enhanced agreement with the computer's conclusion (Figure 26). An increase in performance when SBS-C is used together with explanation is expected as an effect of explanation. The effect of explanation can be directly measured as the improvement in user performance (see Figure 26). Since this is an initial study, we wanted to create a set of scenarios that would help us to detect the effect of explanation, if any, rather than to estimate the correct magnitude of the effect of explanation in real situations. We can detect the effect of explanation in this way only if the following conditions are satisfied:

1. There must be no significant difference between baseline performance of scenarios in the control set of the participant-scenario pairs SPS_C (scenarios evaluated in control mode in the follow-up evaluation) and scenarios in the intervention set of participant-scenario pairs SPS_I (scenarios evaluated in intervention mode).
2. Scenarios must be difficult enough that user performance could benefit from the use of SBS-C.

Assumption 1 allows us to conclude that differences in performance between the control set of participant-scenario pairs SPS_C (evaluated using output of SBS only) and the

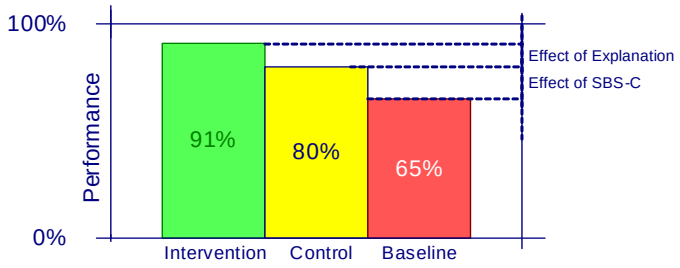


Figure 26: Hypothetical example of the effect of explanation in the case when correct advice is provided by a system

intervention set of participant-scenario pairs SPS_I (evaluated using output of SBS and explanation) are due to explanation rather than to differences between the sets. Assumption 2 allows us to measure the effect of explanation as the difference between the control set of participant-scenario pairs SPS_C and the intervention set of participant-scenario pairs SPS_I in terms of improvements in the participant's accuracy assessment of outbreak scenarios (Figure 26). This reasoning, however, assumes that the participants will be confident enough not to accept the conclusion provided by SBS-C uncritically, without consulting the explanation. If users always accepted the system's advice and the system's advice were always correct, there would be no difference between the control mode (no explanation) and the intervention mode (with explanation), and therefore no room for improvement in user performance. This is why participants evaluated two scenarios in which the system's advice was erroneous regarding the outbreak and the features of the outbreak. Specifically, SBS-B would provide an incorrect conclusion for the two selected scenarios for which HEM provided an explanation. I asked participants to check whether the output of SBS was consistent with the omniscient domain knowledge provided. This process also allowed me to test whether explanation helps the user to recognize incorrect advice provided by the system.

5.5 COMPUTER PRESENTATION

The computer program works in three presentation modes: baseline (displays evidence as counts of patients, see Figure 48), control (information from baseline mode and probabilities calculated using SBS, see Figure 49) and intervention mode (information from control mode and explanation of probabilities calculated using SBS, see Figure 50 and 51). In order to avoid any confounding influence of the program's GUI interface as a result of differences between the control and intervention modes, the interfaces of both modes are as close as possible. In the baseline and both follow up modes the evidence is presented to users, in the form of a graph displaying a time-series of daily counts of monitored people with selected findings (Figure 48). In both the control and intervention modes the posterior probabilities of the nodes in the common part of SBN (Figure 27) are presented to the user (Figure 49). SBN-B (Simple Bayesian Network-Biased) has the same structure as SBN. However, the probability distribution of nodes in the agent subnetworks and nodes in the common part of SBN-B were modified in order to provide incorrect posterior probabilities for these nodes (i.e., the correct outbreak disease does not have the highest posterior probability if calculated using SBN-B). The intervention mode provides an explanation for the results as well (Section 4.2.8). In the control mode, access to the explanation facility is disabled.

The implementation of HEM in the experimental evaluation first presents evidence that decreases the posterior odds of an outbreak disease before including evidence that increases the odds (Section 4.2.5 and Expression 4.18), in order to calculate changes in the posterior odds of the outbreak disease due to the inclusion of evidence groups in the evidence set. Although this ordering is technically correct, the ordering of evidence groups may affect the calculated impact of the current evidence group on the posterior odds of the interface node. I did not want to complicate user interface for participants by allowing them to choose between worst-first and best-first approach.

An alternative approach would be to present supportive evidence first. Either approach yields a display of odds that are technically correct. Future implementation could allow the user to choose the preferred ordering. The impact of any given piece of evidence (in terms of changing the posterior odds) depends on the evidence already displayed to the left of it.

Thus its impact can vary depending on where it is displayed in the evidence ordering. [1]

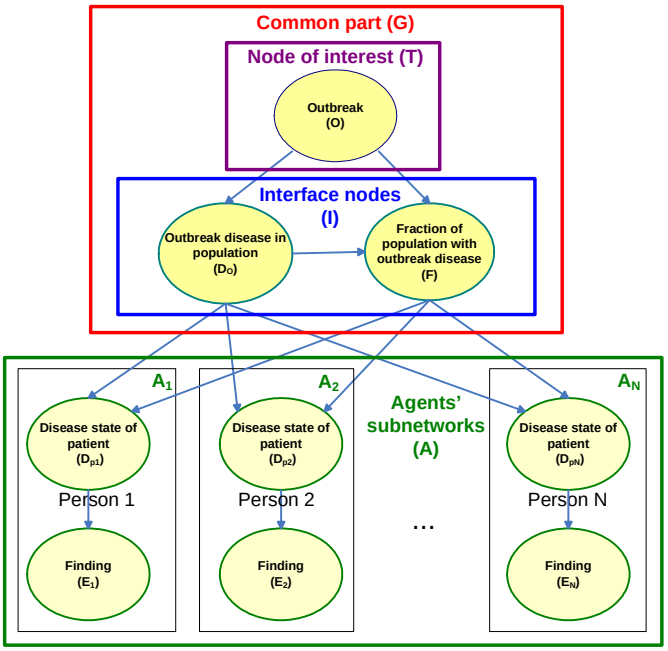


Figure 27: SBN partitioned into subnetworks

I asked participants to read an information sheet about the study, the tasks they were to perform, and probability distributions of the diseases and findings (Table 6). I provided participants with probability distributions of the outbreak (Table 4), diseases (Table 5), and findings in order to familiarize them with the domain represented by SBN. I gave them a brief demonstration of how to use the systems to evaluate outbreak scenarios. In order to increase their learning speed, the available functionality was restricted to what was needed for the study. This domain knowledge was available to the participants during the experiment. During the tutorial, the users were asked to evaluate one scenario using SBS-C in order to familiarize themselves with the domain.

Since, as I mentioned in Section (5.4), incorrect advice is provided by the system for some scenarios, the users were asked to check whether the advice provided by the system was consistent with the domain knowledge.

[1] It would be an interesting future extension to perform a robustness analysis of each piece of evidence over all the locations in the ordering in which it can appear. In a given case, for example, the system might be able to state that evidence "cough" changes the posterior odds by at least 6-fold, regardless of where it is displayed in the evidence ordering.

Table 4: Omniscient probability distribution of outbreak.

Outbreak	Probability
present	0.05
absent	0.95

5.6 STUDY DESIGN

In this section I will describe the design of the study and controls for potential biases. The smallest measurement unit of the study is the participant-scenario pair that represents the participant assessing the scenario. This study follows a scenario-by-scenario two-pass design. Figure 28 summarizes the setting for the quantitative part of this experimental study.

Twenty-two participants took part in the final evaluation of HEM. The participants were asked to provide background information regarding their knowledge of biosurveillance and probability (Figure 46). Based on this assessment, participants were matched pairwise for knowledge of probability and biosurveillance. The purpose of this matching was to create control and intervention sets of participant-scenario pairs with similar performance for the study tasks. To create the participant-participant pairs, I used the collected information about the pool of prospective participants to specify criteria that would result in the best-matched participant-participant pairs. The information about the participants collected using the screening questionnaire consisted of 1) their program of study 2) courses taken in statistics and probability (e.g., none at graduate level, at least one at graduate level), 3) familiarity with terms used in probability, statistics, and epidemiology, and 4) answers to simple questions to verify that they were able to understand the omniscient domain knowledge provided. The participants in eachpair (P_k^1, P_k^2) were randomly split into two sets in order to obtain two randomly matched disjoint sets of participants SP_1 and SP_2 (Figure 29).

The scenarios were split into two sets as well. In order to create two random disjoint

Table 5: Omniscient probability distribution of outbreak disease given that an outbreak has occurred.

Outbreak disease	Probability
disease A	0.05
disease B	0.84
disease C	0.11

matched sets of scenarios, I used the same technique that I used to create sets of participants. First, the scenarios were matched pairwise for difficulty. Since the matching was done in an artificial domain, an expert could not be used to do the matching. I matched scenarios automatically based on the posterior probabilities of there being an outbreak. If the probabilities for two scenarios are sufficiently close (within ±12% range for probabilities > 0.1, within ±56% range for probabilities> 0.01 and$\leqq 0.1$, and within ±72% range for probabilities< 0.01), I considered them to be equally difficult. After the pairs of scenarios(S_l^1, S_l^2) were created, scenarios in each pair were split and randomly assigned to two matched disjoint sets of scenarios SS_1 and SS_2 (Figure 30). Then I created four (2x2) disjoint participant-scenario sets as a Cartesian product $SPS_{i,j} = SP_i \times SS_j$ containing participant-scenario pairs (P, S), where $P \in SP_i$, $S \in SS_j$, $i = 1, 2$, and $j = 1, 2$ (Table 7). After the four sets of participant-scenarios were created they were reorganized into two sets of participant-scenario pairs: an intervention set of participant-scenario pairs SPS_I and a control set of participant-scenario pairs SPS_C. I used the following procedure to assign sets of participant-scenario (PS) pairs $SPS_{i,j}$ into SPS_C and SPS_I: (1) $SPS_{1,1}$ was assigned randomly to the control or intervention set of participant-scenario pairs, (2) $SPS_{2,2}$ was assigned to the same set as $SPS_{1,1}$, and $SPS_{1,2}$ and $SPS_{2,1}$ were assigned to the other set than $SPS_{1,1}$ (Tables 8 and 9). An overview of the steps that were used to construct the control and intervention datasets of participant-scenario pairs is shown in Figure 31. Each participant saw each scenario in the baseline mode and in only one follow-up mode: half of

Table 6: Omniscient conditional probability of chief complaint findings given patient diseases.

Chief Complaint \ Disease	disease A	disease B	disease C	Non-outbreak disease
abdominal pain	0.205	0.000	0.095	0.055
confusion/altered mental status	0.000	0.002	0.466	0.008
cough	0.056	0.336	0.000	0.025
diarrhea	0.264	0.025	0.033	0.007
fever/chills	0.170	0.412	0.121	0.032
other outbreak findings	0.304	0.225	0.285	0.249
other findings	0.000	0.000	0.000	0.624

the scenarios in the "control mode" and half in the "intervention mode". Although each of the sets SPS_C and SPS_I contains all participants and all scenarios, they each contain only half of the unique participant-scenario pairs.

For each pair (P_i, S_j) in SPS_C the participant P_i will evaluate the scenario S_j in the control mode and for each pair (P_i, S_j) in SPS_I the participant P_i will evaluate the scenario S_j in the intervention mode.

The sequence of tasks performed by one participant during the study is shown in the diagram in Figure 32. Each scenario was evaluated by a participant in baseline mode first, and after all scenarios had been evaluated in the baseline mode they were evaluated in follow-up mode (control or intervention mode). The order in which the scenarios were seen by participants was randomized in order to avoid a possible bias due to experience the participants may have acquired from previous scenarios, as judgment about SBS may be influenced by previous experience with the use of SBS with the explanation facility, and similarly, previous experience with use of the SBS may influence judgment about the

Study participants:

18 graduate students, 3 undergraduate students and, 1 faculty participated in the final evaluation.

Scenarios:

Simulated patient records obtained from the emergency department during outbreak (14) and non-outbreak (2) periods.

Modes in which scenarios are presented to participants:

Baseline: Only evidence observed for monitored population.

Control: Only evidence and posterior probabilities from SBS are available.

Intervention: Evidence, posterior probabilities from SBS and HEM explanations are available.

Randomization:

1. Participant-scenario pairs were randomly partitioned into control and intervention sets.

2. Ordering in which the scenario pairs were presented to the user were randomized.

3. Ordering of control and intervention modes for case pair were randomized.

Evaluation task of participants:

Baseline assessment: Interpreting scenarios without SBS using only observed evidence.

Follow-up assessment: Participant assessment of the scenarios using SBS with or without HEM explanation and rating of information provided by the computer.

Figure 28: Summary of quantitative evaluation of hierarchical explanation method for agent-based BNs.

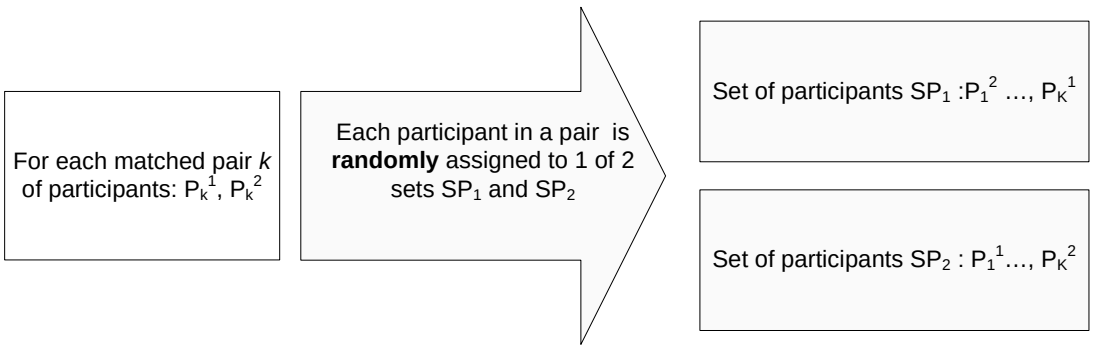


Figure 29: Assigning participants to sets

explanation facility. In this way we can limit the influence of the ordering in which PSs are evaluated.

The participants were asked to provide a baseline and follow-up assessment for each scenario. The follow-up was either in the control or the intervention mode. A questionnaire was used to record the users' assessments. The participants had to complete each scenario before moving on to the next scenario. First, the biosurveillance data (showing daily counts of patients with chief complaint findings) were presented to the user. After the participants had spent sufficient time analyzing the scenario, they were asked to provide a baseline assessment of population health (the possibility of outbreak and outbreak disease). Next, depending on the presentation mode, participants analyzed the case using SBS (control mode) or using SBS with the explanation facility (intervention mode). In the control mode, SBS provides posterior probabilities for the nodes in the common part of the SBN (Figure 27). The only exception is the *"Fraction of Population"* variable, for which I provided the expected number of patients calculated from the posterior distribution of the variable. In the intervention mode both the posterior probabilities of various global nodes and an explanation of these probabilities were available to the participants. After the participants had spent sufficient time studying a scenario in the control or intervention mode, they were again asked to give an assessment of the scenario by means of a follow-up questionnaire. The content of the questionnaire is discussed in the next section (Section 5.7).

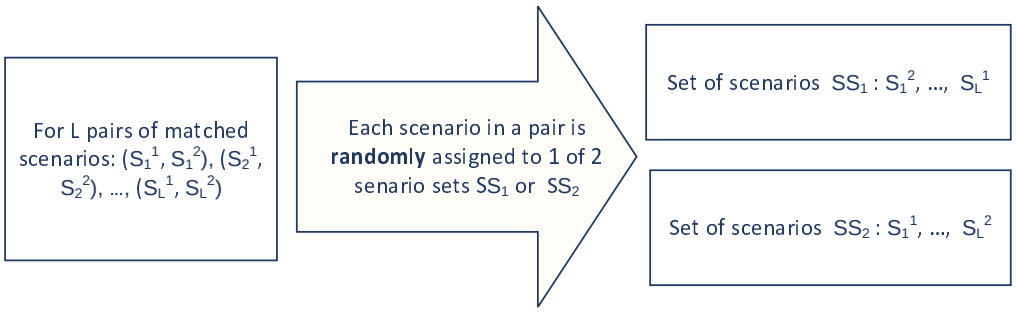


Figure 30: Assigning scenarios to sets

5.7 CONTENT OF QUESTIONNAIRE

Before I asked participants to answer the questions in the baseline questionnaire, I provided the participants a summary of patients' records from an emergency department (ED) (baseline presentation mode). This summary consisted of daily counts of patients coming to the ED with various selected findings. In the baseline questionnaire I asked the questions (Figure 41) that elicited an initial assessment of the situation by a participant. First, a participant was asked to provide his or her probability assessment of the probability of an outbreak. Second, the participant was asked to identify what is the most probable outbreak disease and fractions of the infected population. Third, the participant was asked to provide evidence that he or she would use to support his or her assessment. Participants were asked to answer how confident they were about their answers to each question above.

Before I asked the participants to complete the follow-up questionnaire, I presented them information about the scenario in control or intervention modes. In control mode, in addition to the information that was provided to the participants in baseline mode, they also were given the posterior probabilities of the nodes in a common part of SBN. In the intervention mode, in addition to the information provided to the participants in control mode, they were also given an explanation. In the follow-up questionnaire (Figure 42), participants first filled out for each scenario their assessment of the disease outbreak by answering the same questions as in the baseline questionnaire (Figure 41). In addition, participants were asked whether

Table 7: Creating sets of participant-scenarios pairs.

	Set of scenarios 1 (SS_1)	**Set of scenarios 2** (SS_2)
Set of participants 1 (SP_1)	$SPS_{1,1}$	$SPS_{1,2}$
Set of participants 2 (SP_2)	$SPS_{2,1}$	$SPS_{2,2}$

or not the computer-generated output was consistent with the provided domain knowledge. The participants were also asked to provide his or her opinion about the helpfulness of the specific computer output.

At the end, after participants had seen all scenarios, they were asked to rate various features of SBS and the explanation facility (Figure 43) and provide his or her opinion about the helpfulness of the specific computer output used to answer the questions about the outbreak.

5.8 STATISTICAL ANALYSIS OF THE RESULTS

I analyzed answers provided by participants focusing on 4 categories: scenario classification (probability of an outbreak being present), diagnosis of the outbreak disease, number of people with the outbreak disease, and confidence about an assessment of the probability of an outbreak being present.

The unit of measurement is *the participant-scenario* pair (PS), which represents a scenario evaluated by one study participant. While every scenario belongs to both the control (SPS_C) and the intervention (SPS_I) PS sets, every PS belongs to only the one of these two sets. The PSs in SPS_C and SPS_I sets were evaluated differently in follow-up assessments and equally in the baseline mode. In order to be able to evaluate the effect of different treatments of scenarios in the control and intervention modes, the PSs in SPS_C have to be on average the same (measured by the values of responding variables) as PSs in SPS_I if they are evaluated

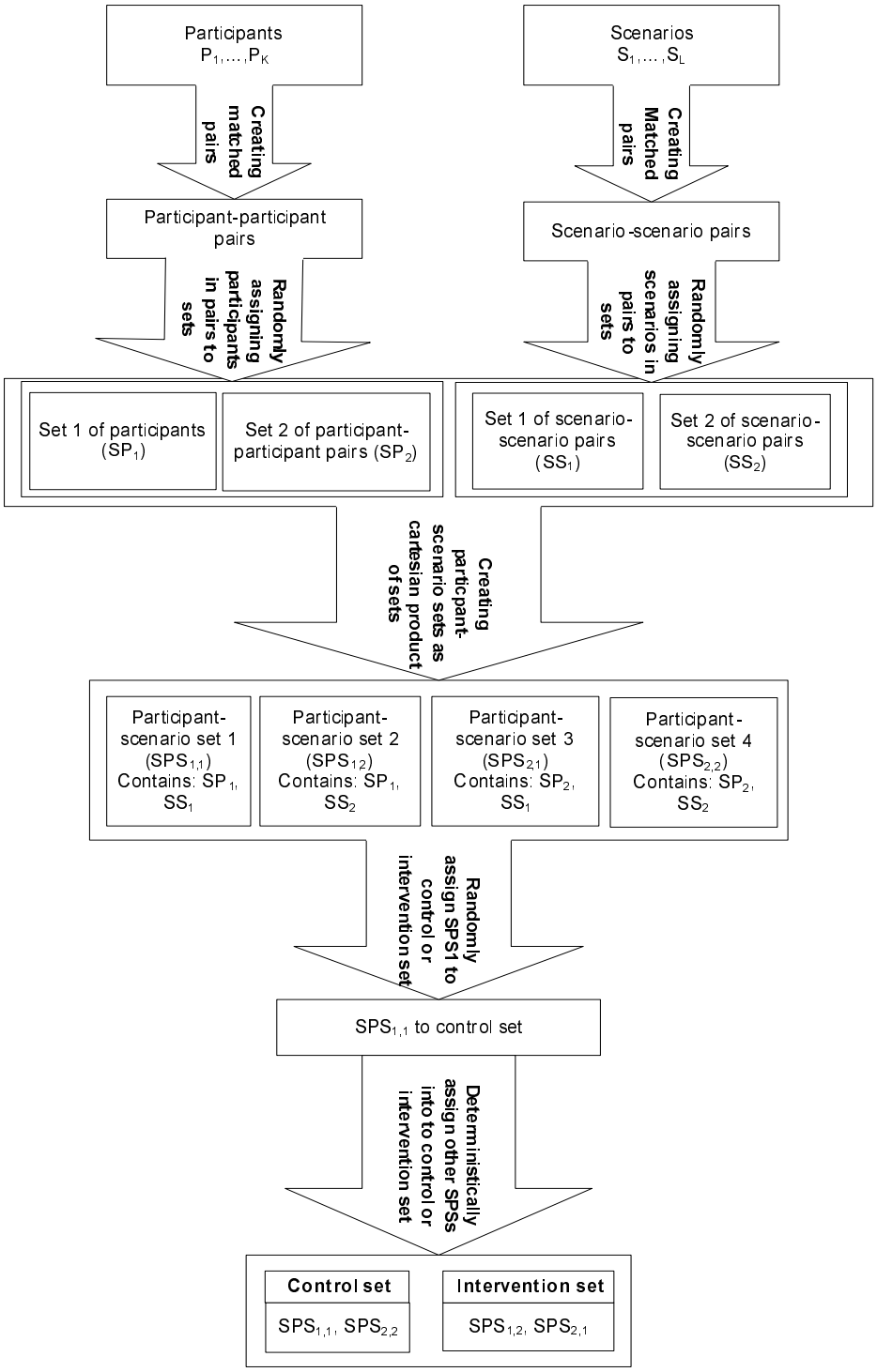


Figure 31: Overview of steps used to construct control and intervention sets.

Table 8: An example of assigning the participant scenarios to the control and intervention sets of participant-scenario pairs.

	Set of scenarios 1 (SS_1)	**Set of scenarios 2** (SS_2)
Set of participants 1 (SP_1)	$SPS_{1,1}$ *to control set*	$SPS_{1,2}$ *to intervention set*
Set of participants 2 (SP_2)	$SPS_{2,1}$ *to intervention set*	$SPS_{2,2}$ *to control set*

in the baseline mode. In the analysis of the results, I compared results for PSs in SPS_C with those PSs as in SPS_I in baseline and follow up modes. Since I had 22 participants in the final evaluation, and each participant saw 8 scenarios in the control mode and 8 scenarios in the intervention mode, I have 176 PSs in the SPS_C and SPS_I. SPS_C and SPS_I were constructed to be matched to the experience of the participants and difficulty of the scenarios to insure that they were equal for the purpose of the evaluation.

Let the symbol X denote some assessment statistic. Then, a superscript will indicate either a baseline X^B or a follow-up assessment X^F, while a subscript will indicate either a control X_C or an intervention X_I set of **PSs**.

Each piece of data collected from the study was obtained from a PS pair. Cases were sampled randomly using the SBS-C model by choosing outbreak disease and the number of people with that disease. Then, I selected scenarios such that they represented various stages of the outbreak. Difficulty to assess a scenario and disease were systematically chosen such that they include all combinations of two diseases that are correctly modeled in SBS-I and 3 levels of difficulty of the scenarios according to the Table 10. The number of patients were obtained from the epidemic model as is described in Section 5.4. Even though the patients and their findings were randomly sampled according to a SBS-C model and selected outbreak disease, the scenarios were correlated in terms of the outbreak disease and strength of the outbreak (difficulty to classify an outbreak correctly). An analysis that is based on the individual observations without taking in to account this clustering is likely to overestimate the statistical significance of any observed effects. Thus, I did not apply traditional

Table 9: Example of control set of participant-scenario pairs, intervention set of participant-scenario pairs, baseline with control set, and baseline with interventions set from Table 8.

Sets for evaluation modes	
Control mode set SPS_C :	$SPS_{1,1}$, $SPS_{2,2}$
Intervention mode set SPS_I:	$SPS_{1,2}$, $SPS_{2,1}$
Baseline mode with control set SPS_{B_C}	$SPS_{1,1}$, $SPS_{2,2}$
Baseline mode with intervention set SPS_{B_I}	$SPS_{1,2}$, $SPS_{2,1}$

MANOVA analysis to test the hypothesis in the Section 1.1. I instead adopted a linear and logistic mixed effects model (West et al., 2006) that takes into account (1) the variations within groups and (2) the correlations between factors disease and difficulty to classify the scenario correctly.

I assigned three factors "disease", "difficulty to classify the scenario correctly", and "level of computer based support (LCBS)" as fixed effects and random effect for this study are participants since they were randomly assigned to SPSs and the outbreak scenarios under specific experimental configurations (disease and difficulty level) are assumed to be independently created as randomly generated patient data.

In Section 5.8.1 I will describe the general approach to analyzing the results.

5.8.1 Normal response variable

This section describes how I evaluated assessments and opinions provided by participants and their derivations that may be assumed to be approximately normally distributed. I will refer to these measurements as response (dependent) variables. Each assessment or opinion was collected in one of the four circumstances: the baseline mode using a control set of SPSs, the baseline mode using an intervention set of SPSs, the control mode using a control set of SPSs, and the intervention mode using intervention set of SPSs.

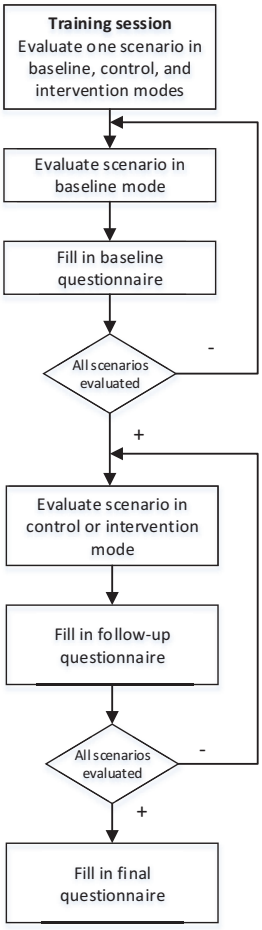


Figure 32: Sequence of activities preformed by one participant during the study

Table 10: Parameters of the scenarios

scenario pair	outbreak disease	Difficulty to classify (based on the strength of the outbreak)
1	cryptopsoridiosis	High
2	cryptopsoridiosis	Medium
3	cryptopsoridiosis	Medium
4	influenza	High
5	influenza	Medium
6	influenza	Low
7	hepatitis	Low
8	no outbreak disease	Low

Version with 2 levels of computer based support In general for normal responding variable Y_{ijk} I used the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{t(i,j)} + \theta_{d(j)} + \zeta_{c(j)} + \epsilon_{ij}, \tag{5.1}$$

where $e_{i,j}$ is an error term and Y_{ij} is a normally distributed variable for Participant i and Scenario j with participant and scenario factors α_i and β_j, respectively, modeled as random effects and LCBS factor $\gamma_{t(i,j)}$ modeled as fixed factor where $t\,(i,j)$ is LCBS with values

$$t\,(i,j) = \begin{cases} 3-1 & \text{control mode compared to a baseline mode with a control set} \\ 4-2 & \text{intervention mode compared to a baseline mode} \\ & \text{with an intervention set} \end{cases}, \tag{5.2}$$

Table 11: Difficulty to classify an outbreak scenario

	Difficulty to classify the scenario as outbreak present/absent (based on the model's probability of the outbreak)			
outbreak disease	High	Medium	Low	Pairs Total
cryptosporidiosis	1 pair	2 pairs	-	3
influenza	1 pair	1 pair	1 pair	3
hepatitis	-	-	1 pair	1
no outbreak disease	-	-	1 pair	1

$\theta_{d(j)}$ is a fixed factor for outbreak disease where $d\,(j)$ is disease with values

$$d\,(j) = \begin{cases} a & \text{disease A} \\ b & \text{disease B} \\ c & \text{disease C} \\ n & \text{no outbreak disease} \end{cases}, \tag{5.3}$$

and $\zeta_{c(j)}$ is a fixed factor 'difficulty to assess an scenario correctly' (DtASC) where $c\,(j)$ is DtASC with values

$$c\,(j) = \begin{cases} H & \text{high - scenario is very difficult to assess correctly} \\ M & \text{medium -scenario is moderately difficult to assess correctly} \\ L & \text{low - scenario it is easy to assess } \textit{correctly} \end{cases}. \tag{5.4}$$

I am using a model with random effects since the data Y_{ij} are not independent. There were 11 data points obtained from the same scenario for each presentation mode (control

Table 12: Categories of difficulty to detect an outbreak based on the strength of outbreak

Difficulty to classify scenario correctly	Range of absolute differences between the omniscient probability of outbreak (1 or 0) and probability calculated using SBS-C model
High (2 pairs)	$\langle 0.85, 1 \rangle$
Medium (3 pairs)	$\langle 0.33, 0.63 \rangle$
Low (3 pairs)	$\langle 0.002, 0.18 \rangle$ (3 pairs)

or intervention) and $n_s/2$ data points obtained from scenarios which were evaluated by the same participant for each presentation mode.

I expected a positive effect of an explanation on the responding variable I evaluated the effect of the explanation on the variable $Y_{i,j}$ by testing a hypothesis

$$H_0 : [(\gamma_{4-2} - \gamma_{3-1})] \leqq 0 \tag{5.5}$$

in contrast to the alternative hypothesis

$$H_1 : [(\gamma_{4-2} - \gamma_{3-1})] > 0 \tag{5.6}$$

with a significance level of test 5%.

Similarly, if I expected a negative effect of an explanation on the responding variable I evaluated the effect of the explanation on the variable $Y_{i,j}$ by testing a hypothesis

$$H_0 : [(\gamma_{4-2} - \gamma_{3-1})] \geqq 0 \tag{5.7}$$

in contrast to the alternative hypothesis

$$H_1 : [(\gamma_{4-2} - \gamma_{3-1})] < 0 \tag{5.8}$$

with a significance level of test 5%.

Version with 4 treatments This model is similar to the model for 2 treatments in Section 5.8.1 where

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{t(i,j,k)} + \theta_{d(j)} + \zeta_{c(j)} + \epsilon_{ijk}, \tag{5.9}$$

where α_i , β_j, θ_j, and ζ_j are the same factors as in the model in Equation 5.1, k is the mode and $t\,(i,j,k)$ is a treatment

$$t\,(i,j,k) = \begin{cases} 1 & \text{baseline mode and control set} \\ 2 & \text{baseline mode and intervention set} \\ 3 & \text{control mode and control set} \\ 4 & \text{intervention mode and intervention set} \end{cases} . \tag{5.10}$$

In general, I evaluated the positive effect of explanation on variable $Y_{i,j,k}$ by testing hypothesis

$$H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \leq 0 \tag{5.11}$$

in contrast to the alternative hypothesis

$$H_1 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] > 0. \tag{5.12}$$

Similarly, I evaluated the negative effect of explanation on variable $Y_{i,j,k}$ by testing hypothesis

$$H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \geq 0 \tag{5.13}$$

against the alternative hypothesis

$$H_1 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] < 0. \tag{5.14}$$

5.8.2 Binomial response variable

In the case of binomial variable the model is given by an expression

$$Y_{ijk} = E\{Y_{ijk}\} + e_{ijk}, \tag{5.15}$$

where $e_{i,j,k}$ is an error term and Y_{ijk} is a Bernoulli response variable for participant i, scenario j and mode k with values

$$Y_{ijk} = \begin{cases} 1 \\ \\ 0 \end{cases}, \tag{5.16}$$

where

$$k \;=\; \begin{cases} 1 & \text{baseline mode} \\ \\ 2 & \text{follow-up mode (control or intervention)} \end{cases}.$$

An expected value $E\{Y_{ijk}\}$ is equal to the probability $P(Y_{ijk} = 1)$, where:

$$E\{Y_{ijk}\} = \frac{exp\left(\mu + \alpha_i + \beta_j + \gamma_{t(i,j,k)} + +\theta_{d(j)} + \zeta_{c(j)}\right)}{1 + exp\left(\mu + \alpha_i + \beta_j + \gamma_{t(i,j,k)} + +\theta_{d(j)} + \zeta_{c(j)}\right)}, \tag{5.17}$$

where α_i , β_j, θ_j, and ζ_j are the same factors as in the model in Subsection 5.8.1, k is the mode and $t(i,j,k)$ is a treatment with values shown in the expression 5.10.

I will evaluate the positive effect of explanation on the responding variable Y_{ijk} by testing the hypothesis

$$H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \leq 0 \tag{5.18}$$

in contrast to the alternative hypothesis

$$H_1 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] > 0$$

with a significance level of test 5%.

Similarly, if I expected a negative effect of explanation on the responding variable Y_{ijk} I tested the hypothesis

$$H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \geq 0 \tag{5.19}$$

in contrast to the alternative hypothesis

$$H_1 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] < 0$$

with a significance level of test 5%.

5.8.3 Category "scenario classification"

In the **category "scenario classification"** I studied the effect of computer advice on the user's detection of the outbreak using answers to questions about the user's subjective probability of outbreak $\hat{P}$ in all the questionnaires (question 1 in baseline and follow-up questionnaires). I performed two types of evaluation. First, I used the posterior probability of outbreak $\hat{P}_p$ directly and compared it with the true probability $P_{Omniscient}$ of outbreak being present. The probability P_{True} has only two values: 1 if an outbreak is present or 0 if the outbreak is absent. I calculated absolute error of assessment as the absolute value of the difference of participant's probability of the outbreak and correct probability of the outbreak:

$$D_P^{LCBS} = \left| P_{Omniscient} - \hat{P}_{LCBS} \right|, \tag{5.20}$$

where level of computer based support (LCBS) index can be B, C, or I for probability estimated in baseline (B), control (C) or intervention (I) LCBS. The values of mean absolute errors in Table 13 indicate that the error is higher than what would be achieved by using uniform random guess which would result in a mean error of 0.5. This result might be due to the users anchoring on the prior probability of an outbreak given in (Table 4), even though they were informed that the scenarios in the evaluation may not follow that prior distribution. Such anchoring could lead to an error rate above 0.5, since 14 out of the 16 scenarios that the participants rated contained outbreaks. I calculated change of an error of the follow-up mode with respect to baseline mode for modes C and I as

$$\Delta_p^C = D_p^C - D_p^B$$

and

$$\Delta_p^I = D_p^I - D_p^B$$

I used a linear model given by Equation 5.1 where $Y_{i,j} = \Delta_p^{LCBS}$ in order to fit the data. Average values of D_P^{LCBS} for different modes are in Table 13.

Table 13: Mean absolute error of probability assessment D_P^{LCBS}or various categories of LCBS

Level of Computer Based Support	Mean absolute error
Baseline mode with control set	0.6419063
Baseline mode with intervention set	0.6345250
Control mode with control set	0.5594830
Intervention mode with intervention set	0.5429859

After fitting the data using the linear model with fixed effects model in Equation 5.1, I checked that underlying assumption of the model were satisfied by the data. The plot of standardized residuals versus fitted values from the model shown in Figure 33 does not indicate a violation of the assumption of constant variance. In addition, the normal plot in Figure 34 indicates that assumption of normality of within-group residuals is plausible.

I evaluated the following null hypothesis $H_0 : \gamma_{4,2} - \gamma_{3,1} \geq 0$ which states that coefficient of intervention mode increased an error more or decreased and error less than coefficient of control mode versus alternative hypothesis $H_1 : \gamma_{4,2} - \gamma_{3,1} < 0$. Results shown in Table 14 reveal that we can not reject null hypothesis that $H_0 : \gamma_{4,2} - \gamma_{3,1} \geq 0$ ($p - value = 0.431$).

5.8.4 Category "diagnosis of outbreak disease"

In the **category "diagnosis of outbreak disease"** I studied the effect of the explanation on the user's diagnosis of the outbreak disease (question 1b on baseline and follow-up questionnaires). For each participant-scenario pair we have 3 diagnoses $\widehat{dx_B}$ obtained from baseline assessment $\widehat{dx_F}$ or obtained from one of the follow-up modes, and correct diagnosis dx^T, which was used to sample data for a scenario (see Section 5.4). Increase in the proportion of scenarios with the correctly classified outbreak disease was larger in Control LCBS than in Intervention LCBS, as can be seen in Table 16. I evaluated the effect of explanation

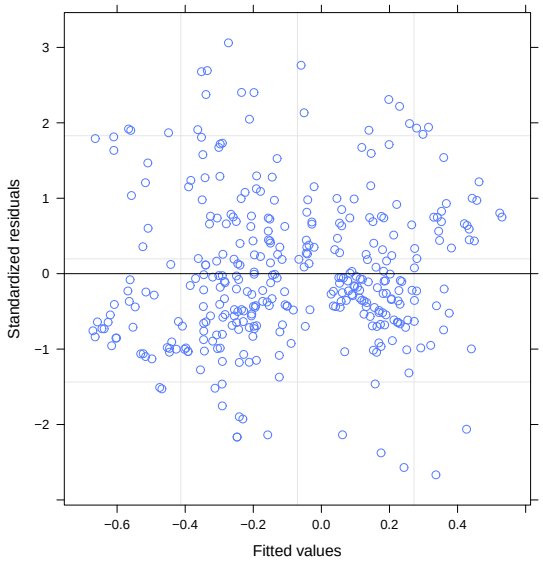


Figure 33: Scatter plot of the standardized within-group residuals versus within-group values for the fitted model

on the diagnosis of the outbreak disease using

$$Y_{ijk} = \begin{cases} 1 & \text{if subject correctly identifies outbreak disease} \\ 0 & \text{otherwise} \end{cases}.$$

I used the logistic regression model described in Section 5.8.2. I have checked the assumptions of the model. A plot of the residuals indicates violation of normality of residuals (Figure 36) since quantiles of the sample do not match theoretical quantiles of normal distribution. Assumption of homogeneity of the variance (Figure 35) is plausible as the variance in the graph does not seem to to depend on fitted values. Possible violation of normality of response variable may cause the model we are using to fail in representing dependencies in the data sufficiently well. If explanation has a positive effect on the classification of an outbreak disease, we expect $[(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \geq 0$. However, we obtained $(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1) = -0.3232$, which suggests that explanation has a negative effect on classification of the outbreak disease. Moreover the p-value obtained from the model for null hypothesis is 0.805. Clearly

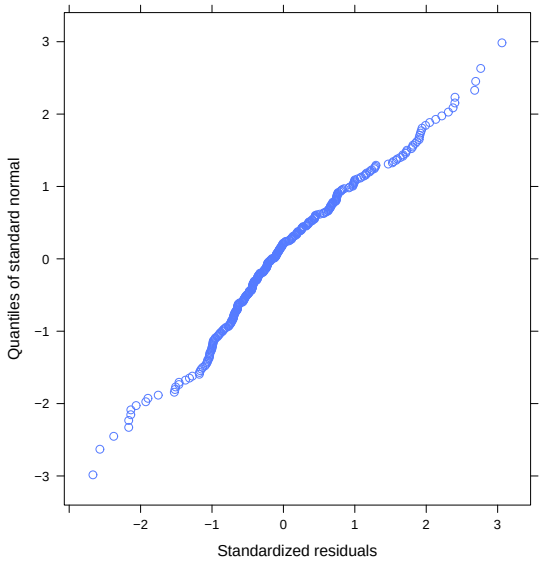


Figure 34: Normal plot of standardized residuals for the fitted model for the assessment of a probability of an outbreak.

we cannot reject null hypothesis $H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \leqq 0$. More information is in Table 15.

We tested the reversed null hypothesis $H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \geq 0$. We also are not be able to rejected this null hypothesis, since the p-value is 0.195.

5.8.5 Category "Number of people with outbreak disease in ED"

In the **category "Number of population with outbreak disease in ED"** I measured the impact of explanation on the users' estimates of the number of people with the outbreak disease (question 3 on baseline and follow-up questionnaires). For each participant-scenario pair we have two estimates of the number of people: N^B from the baseline and N^F from the follow-up assessments, which can be from control N^F_C or intervention mode N^F_I. I used "square root" transformation in order to improve normality of the variable Y_{ij}. The correct number of patients that came to ED due to outbreak disease is N^T. I used the following

Table 14: Result for change of probability of outbreak error in follow-up versus control LCBS given by $\gamma_{4,2} - \gamma_{3,1}$.

Results for change of probability of outbreak error in follow-up versus control LCBS, $Y_{ij} = \Delta_p^F$		Sample size: 352
Null Hypothesis:Coefficient of Intervention mode increased an error more or decreased and error less than coefficient in Control mode if measured versus Baseline LCBS.	$\gamma_{4,2} - \gamma_{3,1}$ (In case of positive effect of an explanation we expect negative number)	p-value
$H_0 : \gamma_{4,2} - \gamma_{3,1} >= 0$	-0.00644	0.431

linear model with mixed effects to fit the data

$$Y_{ij} = \mu + \alpha_i + \beta_j + \theta_j + \zeta_j + \gamma_{t(i,j)} + \epsilon_{ijk}, \tag{5.21}$$

where α_i , β_j, $\theta_{d(j)}$, $\zeta_{c(j)}$ and $\gamma_{t(ij)}$ are the same factors and indices have the same meaning as in the model in Subsection 5.8.1, and

$$Y_{ij} = \Delta^{t(i,j)}$$

where $\Delta^{t(i,j)}$ is given by

$$\Delta^{3,1} = \left|\sqrt{N^T} - \sqrt{N_C^F}\right| - \left|\sqrt{N^T} - \sqrt{N^B}\right|$$

and

$$\Delta^{4,2} = \left|\sqrt{N^T} - \sqrt{N_I^F}\right| - \left|\sqrt{N^T} - \sqrt{N^B}\right|.$$

Table 15: Results for classification of outbreak disease comparing improvement in follow-up versus control LCBS.

Effect of explanation on classification of an outbreak disease		Sample size:
Null hypothesis: Coefficient of an Intervention LCBS increases classification accuracy more than coefficient of Control LCBS if increase is measured relative to coefficient of a Baseline LCBS	$[(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)]$ (In case of positive effect of an explanation we are expecting positive number)	p-value
$H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \leqq 0$	-0.3232	0.805

After fitting the data I checked that the underlying assumptions of the model are satisfied by the data. The plot of standardized residuals versus fitted values from the model shown in Figure 37 indicates plausibility of the assumption of constant variance. In addition, the normal plot in Figure 38 indicates that assumption of normality of within-group residuals is plausible in the interval <-1,1>.

I checked the assumptions of normality and homoscedasticity using a plot of residuals. I tested the null hypothesis that $H_0 : [(\gamma_{4,2} - \gamma_{3,1})] \geq 0$ with an alternative hypothesis $H_0 : [(\gamma_{4,2} - \gamma_{3,1})] < 0$. I can not reject the null hypothesis as the p-value received from fitted linear model for the null hypothesis is 0.14, which is less than the chosen 5% confidence level (see Table 17).

5.8.6 Category "confidence"

In the **category "confidence"** I looked at the subjective rankings of the user's confidence in his or her assessment. Users ranked his or her confidence on a seven-point Likert scale (1-not at all confident to 7-completely confident). I hypothesize that the explanation will increase the confidence of participants in their assessments. I used a linear model

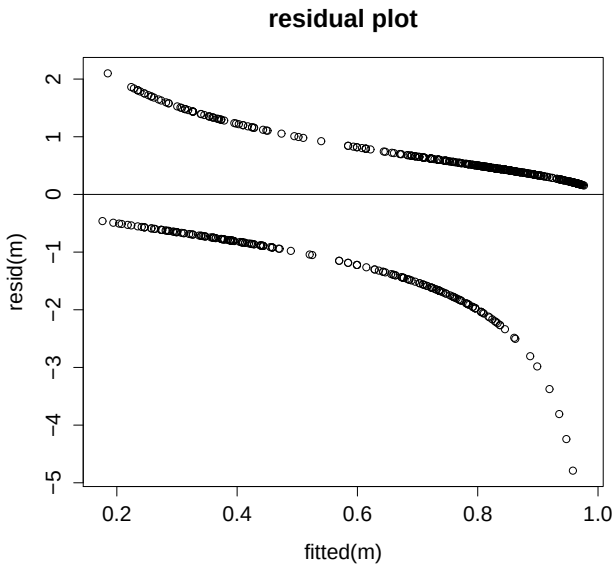


Figure 35: Scatter plot of the standardized within-group residuals versus within-group values for the fitted model for correct assessment of an outbreak disease

with mixed effects where I also transformed scale of confidence ranking such that $Y_{ijk} \in \{-3, -2, \ldots, 0, \ldots, +2, +3\}$. and treated Y_{ijk} as a continuous variable. I used the model in Section 5.8.1 to fit the data. I can not reject null hypothesis $H_0 : [(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \leq 0$ since the calculated p-value for null hypothesis is 0.23. For value of performance measure see Table 18. I checked that the underlying assumptions are reasonably satisfied by the data. The plot of standardized residuals versus fitted values from the model in Figure 39 does not indicate a violation of the assumption of constant variance. In addition, the normal plot in Figure 34 indicates that the assumption of normality of within-group residuals is plausible.

5.8.7 Discussion of Results

The results in all four categories I analyzed were not statistically significant for the hypothesis that explanation has a positive effect on performance in these categories. However, results showed a positive trend in the effect of a explanation in three categories: probability of

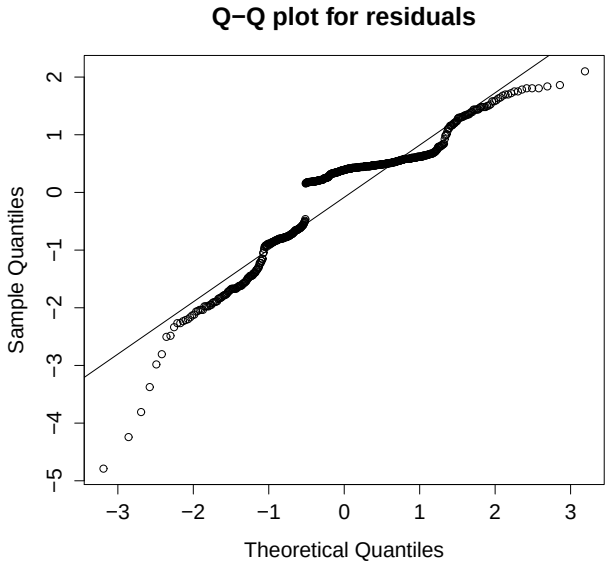


Figure 36: Normal plot of standardized residuals for the fitted model for correct assessment of an outbreak disease

outbreak being present, number of people with outbreak disease and confidence about the assessment of the probability of an outbreak being present. It is possible that with more data the effect of explanation would be statistically significant.

I used a t test to calculate the size of the data sample required to obtain statistically significant results for the category "scenario classification". The number of scenarios required to obtain a statistically significant result if the effect size is taken to be the value of current sample -0.04791302 is 7,461, given significance level 0.05, power level 0.9, and using a one-sided t test. If 16 scenarios are reviewed per participant, the number of participants required is approximately 466, which is clearly a large number and outside of the scope that is feasible for this dissertation.

Table 16: Proportions of scenarios with correctly identified outbreak disease.

LCBS	Proportion of scenarios with correctly identified outbreak disease	Change in proportion in Follow-up LCBS with respect to Baseline LCBS
Baseline, control set	0.625	0
Baseline, intervention set	0.693	0
Control	0.727	0.102
Intervention	0.739	0.46

Table 17: Results for error of assessment in the number of people with the outbreak disease.

Results for error of assessment of number of people with outbreak disease in Follow-up LCBS measured versus Baseline LCBS Δ.		Sample size: 280
Null hypothesis: Explanation does not decrease the difference in error of an assessment more than Control LCBS if difference is measured with respect to Baseline LCBS	$(\gamma_{4,2} - \gamma_{3,1})$ (In case of desired effect of explanation we expect negative number)	p-value
$H_0 : [(\gamma_{4,2} - \gamma_{3,1})] \geq 0$	-0.2911709	0.1051

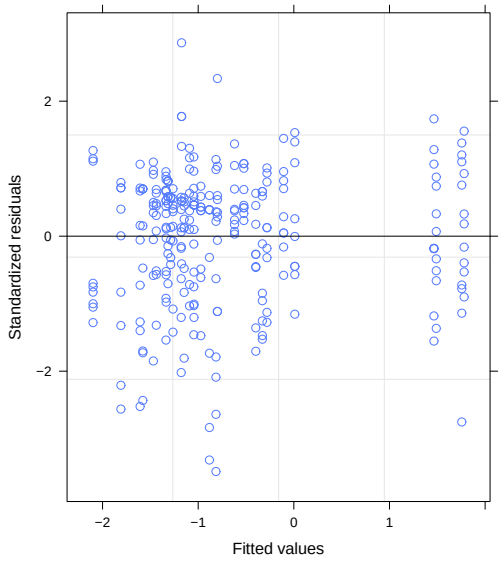


Figure 37: Scatter plot of the standardized within-group residuals versus within-group values for the fitted model (for number of patients with outbreak disease)

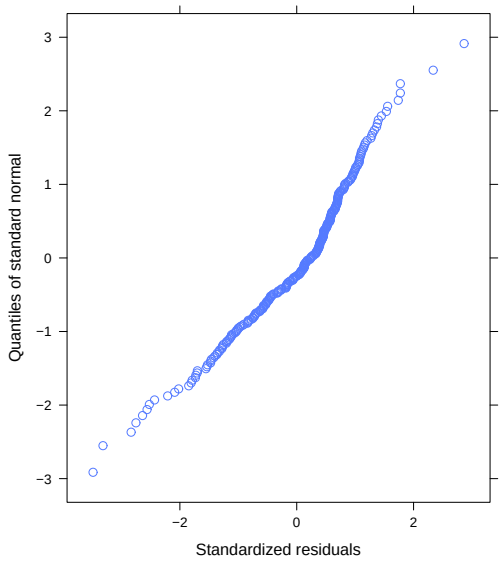


Figure 38: Normal plot of within-group standardized residuals for the fitted model (for number of patients with outbreak disease)

Table 18: Results for confidence in assessment about the probability of an outbreak being present.

Result for confidence of an assessment about a probability of an outbreak being present		Sample size: 624
Null hypothesis: A Coefficient of an Intervention LCBS increases confidence less than coefficient of an Control LCBS (if increase is measured relatively to the coefficient of a Baseline LCBS)	$[(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)]$ (We expect a positive number if explanation increases confidence more than Control LCBS)	p-value
H_0 : $[(\gamma_4 - \gamma_2) - (\gamma_3 - \gamma_1)] \leq 0$	0.1090	0.23

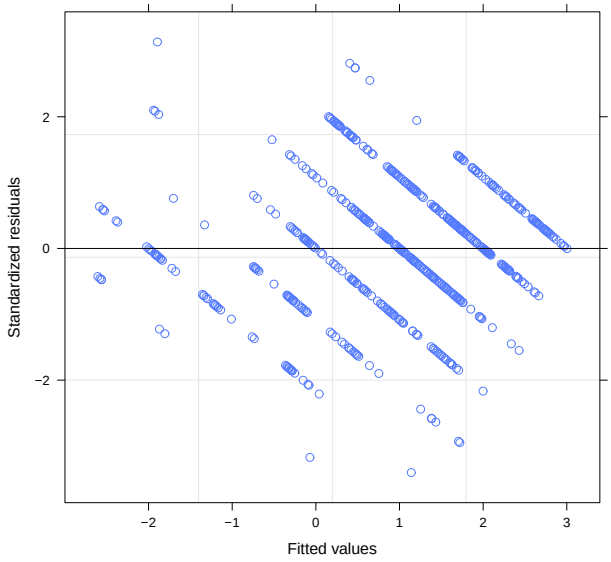


Figure 39: Scatter plot of the standardized within-group residuals versus within-group values for the fitted model for the confidence of the participants about their estimates of probability of an outbreak.

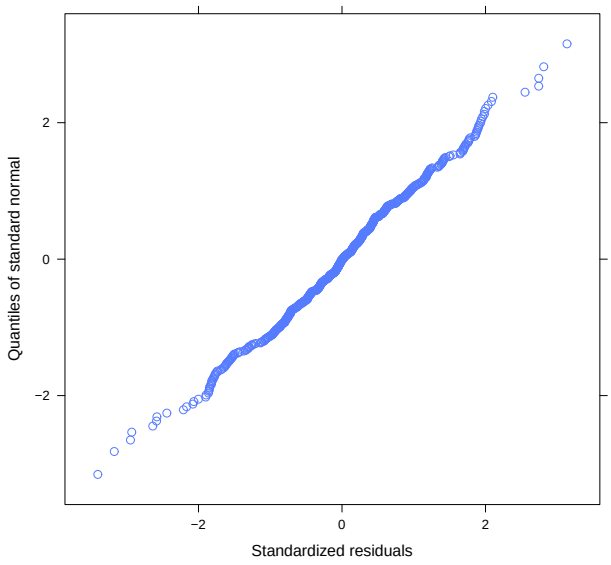


Figure 40: Normal plot of standardized within-group residuals for the fitted model for the confidence of the participants in their estimates of probability of an outbreak.

6.0 CONCLUSIONS AND FUTURE WORK

The results presented in this dissertation do not support the thesis that explanation provided by the proposed explanation method improves users' assessments about the disease outbreaks, or their confidence about their assessments. In this chapter I summarize major contributions of this dissertation and outline topics for future research.

6.1 CONTRIBUTIONS

The main contribution of this thesis is the development and evaluation of methods for explaining the inference on BN that represents a population of independent agents. The HEM is a method to explain inference in BN which models a population of conditionally independent agents, each of which is modeled as a subnetwork. The HEM complements previous approaches that model a single agent (e.g., explaining the diagnostic inference for a single patient using a BN that models that patient). Unlike previous explanation methods, the HEM exploits the modular character of a BN with a population of independent agents. Users of the HEM do not need to be familiar with BNs since, unlike some previous explanation methods, which used all of or a part of BN as presentation of the explanation, the HEM presents explanation independently of the underlying BN using text and diagrams. Like most explanation methods for probabilistic systems, HEM expects that the user has basic knowledge of probability. The lessons learned from the experiments provide directions for improving such explanation systems in the future and identify challenges of evaluating the effect of explanation on decision making.

6.2 FUTURE WORK

Areas of possible future research based on the proposed explanation method include the following:

Additional information and presentation: Responses of study participants indicated that explanation could be improved by including a time dimension in the explanation that would enable users to spot the trends over the time.

User modeling: The level of experience of the user has an impact on the understandability of explanation. Currently, the information and level of detail provided on the screen do not change with the user's experience. This applies not only to various initial experience of the user, but also to experience gained over time. By modeling a user's knowledge explicitly, we may be able to improve the explanation.

Experimental design: Based on experience gained from the experimental study, users focused primarily on completing the task of assessing the outbreaks. The motivation to use explanation or understand recommendation of the system was minimal. A potential solution would be to evaluate the explanation method using experienced public health officials who are well aware of the consequences of an incorrect analysis. They are inherently motivated to understand the results provided by the computer system. Another option would be to find a way to increase motivation of the participants to base their decisions on a deeper understanding of how the probabilities were calculated using the model and the data.

APPENDIX A

QUESTIONNAIRES

Scenario/Case number: ***1*** *Baseline Questionnaire*

The following items assess your initial belief that there is a disease outbreak in for a given case.

			How confident are you about your response?						
			Not at all						Completely
		Response	1	2	3	4	5	6	7
1. According to your assessment, what is the **probability that an outbreak is present** at the end of the presented period?	→	.....	☐	☐	☐	☐	☐	☐	☐
2. Which of the listed **outbreak diseases is most likely** A, B, or C?	→	.....	☐	☐	☐	☐	☐	☐	☐
3. What is your estimate of the **number of people with the outbreak disease** that you list in your answer to question **2** above?	→	.....	☐	☐	☐	☐	☐	☐	☐

4. Imagine, that your manager asks you to justify how you arrived at your conclusions about the probability of an outbreak (1) and about the most likely outbreak disease (2). Explain your answer using data about patients (chief complaints).

Figure 41: Baseline Questionnaire

Scenario/Case number:

Instructions: Please use primarily the **screen "……..."** to assess the disease outbreak and to answer the questions. You may in addition also use other screens if needed.

For each question fill out questionnaire on the helpfulness of the provided information (it is in a separate booklet).

The following items assess your initial belief that there is a disease outbreak for a given case.

			How confident are you about your response?						
			Not at all						Completely
		Response	1	2	3	4	5	6	7
1. According to your assessment, what is the **probability that an outbreak is present** at the end of the presented period?	→	.....	☐	☐	☐	☐	☐	☐	☐
2. Which of the listed **outbreak diseases is most likely** A, B, or C?	→	.....	☐	☐	☐	☐	☐	☐	☐
3. What is your estimate of the **number of people with the outbreak disease** that you list in your answer to question **2** above?	→	.....	☐	☐	☐	☐	☐	☐	☐

4. Imagine, that your manager asks you to justify how you arrived at your conclusions about probability of outbreak (1) and about the most likely outbreak disease (2). Explain your answer using data about patients (chief complaints).

5. Are the computer-generated conclusions consistent with the provided domain knowledge?

no ☐ yes ☐

Figure 42: Follow-up Questionnaire

Final Questionnaire – Fill in only at the end of the user study!

I. Please provide your assessment of how helpful the information on the screens of the biosurveillance system was relative to the screen **"Chief Complaint Findings' Counts"** when making your assessments.

1. Making your assessment of **the probability that there is an outbreak?**

Screen	Much less helpful 1	2	3	4	5	6	Much more helpful 7
Posterior probabilities	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

Screen	1	2	3	4	5	6	7
Explanation	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

2. Making your assessment of what is **the outbreak disease**?

Screen	Much less helpful 1	2	3	4	5	6	Much more helpful 7
Posterior probabilities	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

Screen	1	2	3	4	5	6	7
Explanation	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

Figure 43: Final Questionnaire(page 1)

3. Making your assessment of what is the **number of people with the outbreak disease**?

Screen	Much less helpful 1	2	3	4	5	6	Much more helpful 7
Posterior probabilities	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

Screen	1	2	3	4	5	6	7
Explanation	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

II.

Please provide your assessment of how helpful the information on the screens of the biosurveillance system was when making your **assessments** of the **consistency of the computer-generated conclusions** with the provided domain knowledge?

Screen	Much less helpful 1	2	3	4	5	6	Much more helpful 7
Posterior probabilities	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

Screen	1	2	3	4	5	6	7
Explanation	☐	☐	☐	☐	☐	☐	☐

Please explain why or why not.

Figure 44: Final Questionnaire (page 2)

III. For each of the following features of the biosurveillance system and explanation facility, please provide your assessment of how helpful the feature was when making your assessments of disease outbreaks.

Feature of the system	Not at all helpful 1	2	3	4	5	6	Extremely helpful 7
The computer system's **posterior probabilities** (e.g., the posterior probability of an outbreak).	☐	☐	☐	☐	☐	☐	☐
Textual explanations of how the system derived its posterior probabilities.	☐	☐	☐	☐	☐	☐	☐
Including explanations of states of the variables "*Outbreak Disease" and "Number of People* with *Outbreak Disease"* into the textual explanation.	☐	☐	☐	☐	☐	☐	☐
Graphical explanations of the effect of evidence on the variables *Outbreak Disease* and *Number of People with Outbreak Disease*.	☐	☐	☐	☐	☐	☐	☐
Other:_____ _____ _____	☐	☐	☐	☐	☐	☐	☐

IV. Please provide any suggestions you have for improving automated explanations of computer-assisted outbreak detection.

.....

.....

.....

.....

.....

.....

.....

.....

.....

Figure 45: Final Questionnaire (page 3)

Information required to help create a balanced study

Thank you for your interest in participating in a research study involving "An evaluation of automated explanation of a disease-outbreak detection system". Before you can be enrolled in the research study, I need to ask you some questions that will help me to create a balanced study.

The main purpose of this study is to evaluate the practical value of a computer-based system that automatically detects disease outbreaks and explains its diagnostic reasoning.

I need to acquire some information which will be used solely to create matching pairs of participants who have comparable skills and background.

Answering these questions is completely voluntary, but necessary for participation in the study.

Please fill out the attached questionnaire and email it to sutovsky@gmail.com.

1. Are you studying for a taught MSc (or PgDip or PgCert)? If yes, what is the name of the course?

   _____

   Are you studying for a research degree (PhD, MPhil, MSc by research)? If yes, please state the general area of your research or its working title.

   _____

2. Have you studied any courses in probability and statistics?
   a. at school or college _____
   b. at the undergraduate level _____
   c. at the graduate level (e.g., Introduction to Statistical Methods, Applied Regression Analysis)_____

3. Please use the scale below to rate your familiarity with the following terms:

	None	Have heard of the term but unsure of it (before participating in the study)	Somewhat familiar in the past	Very familiar now
	1	2	3	4
a. Conditional probability	() 1	() 2	() 3	() 4
b. Posterior probability	() 1	() 2	() 3	() 4
c. Bayes' theorem	() 1	() 2	() 3	() 4
d. Prevalence	() 1	() 2	() 3	() 4
e. Likelihood ratio	() 1	() 2	() 3	() 4

Figure 46: Screening questionnaire (part 1)

f. Disease outbreak detection	()	()	()	()
	1	2	3	4

4. Please, try your best to answer the question below. It is not necessary to answer all the questions in order to participate in the study, but I ask you to make at least an attempt to answer them. Provide even a partial answer. If you do not know a solution please write: "I do not know".

   Using the probability distribution of diseases in Table 1 and the conditional probability distributions in Table 2, answer the following questions. You may assume that symptoms are conditionally independent of each other, given a disease.

Disease	Probability
disease a	0.84
disease b	0.05
disease c	0.11

Table 1. Probability distribution that a patient has specific disease i.e.,P(disease)

Disease / Symptom	Disease a	Disease b	Disease c
abdominal pain	0	0.230	0.095
confusion	0.002	0	0.466
cough	0.336	0.063	0
diarrhea	0.025	0.295	0.033
fever/chills	0.412	0.190	0.121
other findings	0.225	0.223	0.285

Table 2. Conditional probability distribution P(symptom|disease)

a) Which 2 symptoms are patients with "disease b" most likely to suffer from?

   Briefly explain your answer.

b) If a patient has a cough and a fever/chills, which disease is the patient most likely to have?

   Briefly explain your answer.

Figure 47: Screening questionnaire (part 2)

APPENDIX B

EXAMPLE OF HEM OUTPUTS

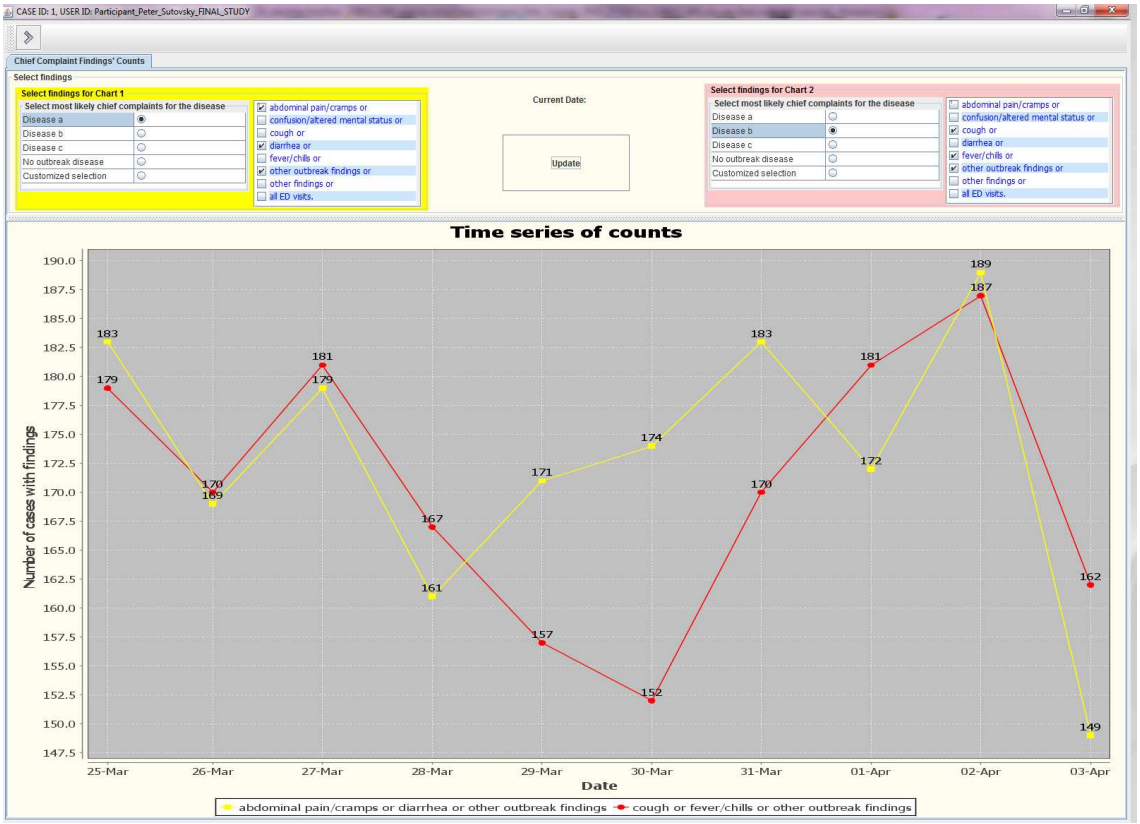


Figure 48: Example of the HEM screen displaying time series of patient counts..

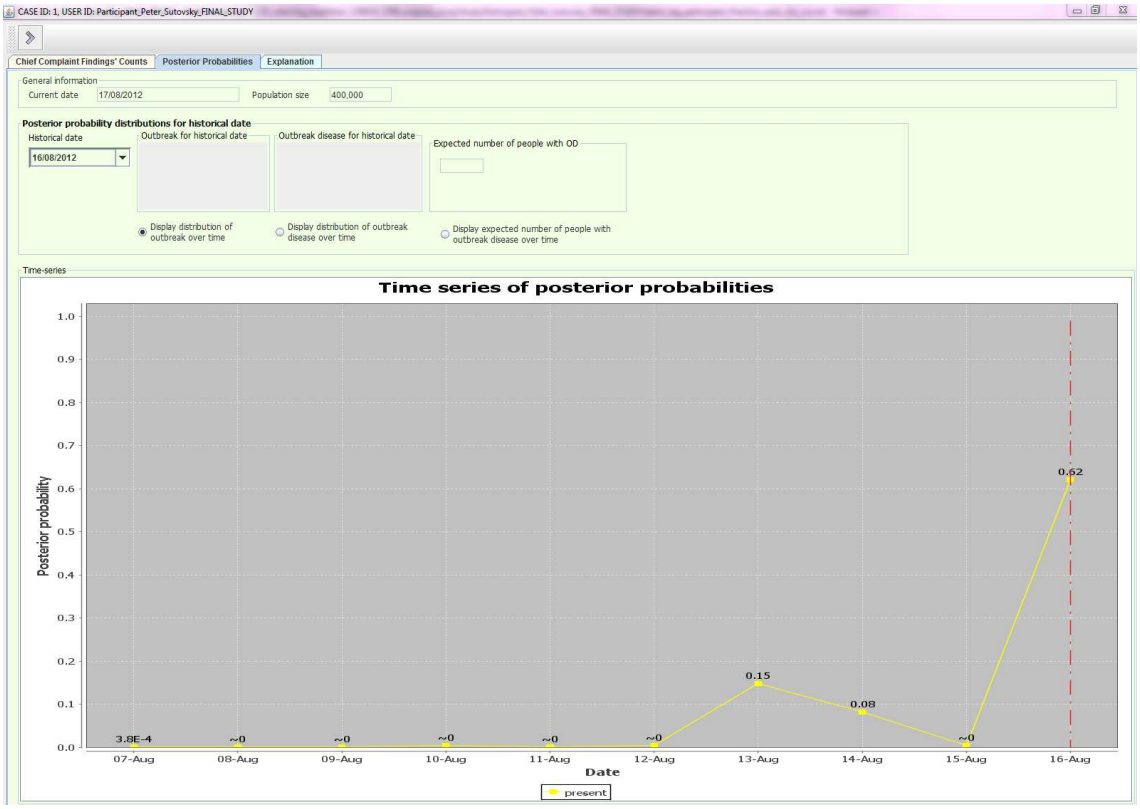


Figure 49: Example of the HEM screen displaying time series of posterior probabilities of common nodes, the main screen of Control LCBS.

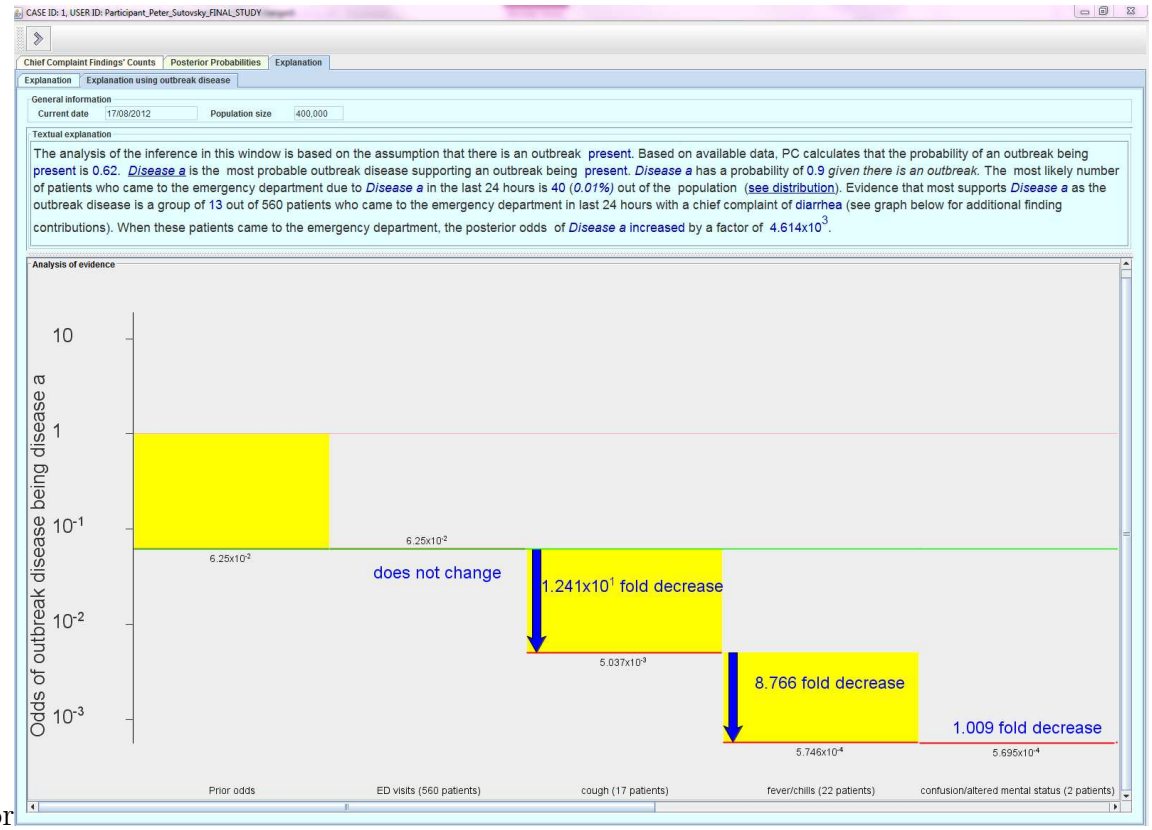
for 

Figure 50: Example of the HEM screen displaying explanation, the main screen of Intervention LCBS (part 1).

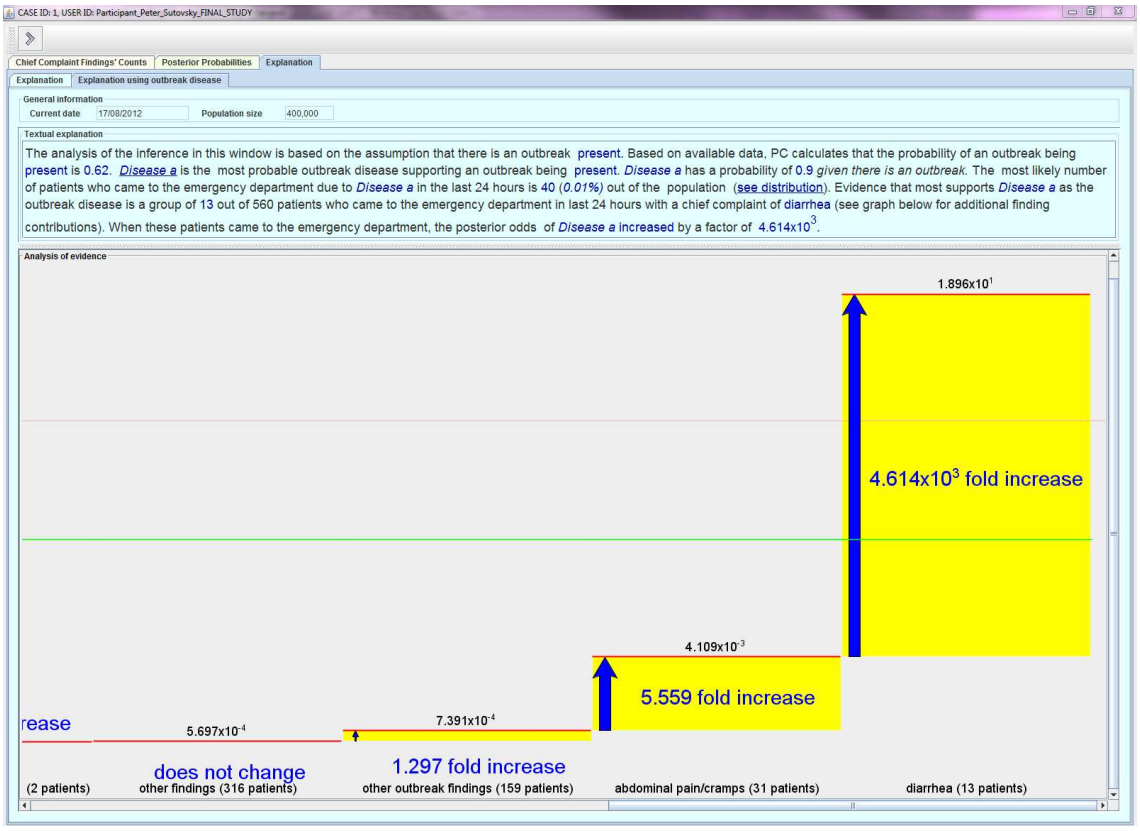


Figure 51: Example of the HEM screen of explanation, the main screen of Intervention LCBS, scrolled horizontally to the right (part 2).

Bibliography

S. K. Andersen, K. G. Olesen, and F. V. Jensen. Hugin–a shell for building Bayesian belief universes for expert systems. pages 332–337, 1990.

M Baker and T Boult. Pruning Bayesian networks for efficient computation. Mountain View, CA, 1990. Association for Uncertainity and Artificial Intelligence.

Bayesia SA. *BayesiaLab*. URL www.bayesia.com.

Dianne C. Berry and Donald E. Broadbent. Explanation and verbalization in a computer-assisted search task. *The Quarterly Journal of Experimental Psychology*, 39a(4):585 – 609, November 1987.

Bruce G. Buchanan and Edward H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1984.

Svetlana Bulashevska, Orsolya Szakacs, Benedikt Brors, Roland Eils, and Gyula Kovacs. Pathways of urothelial cancer progression suggested by Bayesian network analysis of allelotyping data. *International Journal of Cancer*, 110(6):850–856, 2004. URL http://dx.doi.org/10.1002/ijc.20180.

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

B. Chamberlain and T. Nordahl. Conflict detection in causal probabilistic networks. Master's thesis, Institute for Electronic Systems, Aalborg University, Aalborg, Denmark, 1989.

Eugene Charniak. The Bayesian basis of common sense medical diagnosis. In *AAAI*, pages 70–73, 1983.

Cristina Conati, Abigail S. Gertner, Kurt VanLehn, and Marek J. Druzdzel. On-line student modeling for coached problem solving using Bayesian networks. In *Proceedings of the Sixth International Conference on User Modeling (UM-96)*, pages 231–242, Vienna, New York, 1997. Springer Verlag.

Gregory F. Cooper. *NESTOR: A Computer-based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge*. PhD thesis, Stanford University, 1984.

Gregory F. Cooper, John N. Dowling, John D. Levander, and Peter Sutovsky. A Bayesian algorithm for detecting cdc category a outbreak diseases from emergency department chief complaints. In *Advances in Disease Surveillance*. International Society for Disease Surveillance, 2006.

A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247, 1968.

Pedro Domingos and Michael J. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

Marek J. Druzdzel. Qualitative verbal explanations in Bayesian belief networks. *Artificial Intelligence and Simulation of Behaviour Quarterly*, 94:43–54, 1996.

Marek J. Druzdzel. SMILE: Structural modeling, inference, and learning engine and GeNIe: A development environment for graphical decision-theoretic models (intelligent systems demonstration). In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 902–903, Menlo Park, CA, 1999. AAAI Press/The MIT Press.

Marek J. Druzdzel and Max Henrion. Using scenarios to explain probabilistic inference. In *Working notes of the AAAI–90 Workshop on Explanation*, pages 133–141, Boston, MA, 1990. American Association for Artificial, Intelligence.

Marek J. Druzdzel and Max Henrion. Efficient reasoning in qualitative probabilistic networks. In *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93)*, pages 548–553, Menlo Park, CA, 1993. AAAI Press/The MIT Press.

Marek J. Druzdzel and Henri J. Suermondt. Relevance in probabilistic models: "backyards" in a "small world". In *Working notes of the AAAI–1994 Fall Symposium Series: Relevance*, pages 60–63, New Orleans, LA, 4–6 November 1994.

R. O. Duda, P. E. Hart, and N. J. Nilsson. Subjective Bayesian methods for rule-based inference systems. pages 274–281, 1990.

Joseph N. S. Eisenberg, Edmund Y. W. Seto, Jr. Colford, John M., Adam Olivieri, and Robert C. Spear. An analysis of the milwaukee cryptosporidiosis outbreak based on a dynamic model of the infection process. *Epidemiology*, 9(3):255–263, 1998. ISSN 10443983. URL http://www.jstor.org/stable/3703054.

Andre Michael Everett. *An empirical investigation of the effect of variations in expert system explanation presentation on users' acquisition of expertise and perceptions of the system*. PhD thesis, The University of Nebraska - Lincoln, 1994.

B. J. Frey. *Graphical models for machine learning and digital communication*. Cambridge, Massachusetts, London, England, 1998.

Charles P. Friedman, Arthur S. Elstein, Fredric M. Wolf, Gwendolyn C. Murphy, Timothy M. Franz, Paul S. Heckerling, Paul L. Fine, Thomas M. Miller, and Vijoy Abraham. Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation: A Multisite Study of 2 Systems. *JAMA*, 282(19):1851–1856, 1999. doi: 10.1001/jama.282.19.1851. URL http://jama.ama-assn.org/cgi/content/abstract/282/19/1851.

Robert Fung and Brendan del Favero. Applying Bayesian networks to information retrieval. In *Communications of the ACM*, volume 38, pages 42–48, 1995.

R. W. Gault. *Learning and Explanation Type in a Knowledge-based Arms Control Inspection Assistant: An Empirical Evaluation (Strategic Arms Reduction Treaty)*. PhD thesis, George Washington University, 1994.

Jörg Gebhardt, Aljoscha Klose, Heinz Detmer, Frank Rügheimer, and Rudolf Kruse. Graphical models for industrial planning on complex domains. (482):131–143, 2006. URL http://fuzzy.cs.uni-magdeburg.de/~ruegheim/publications/cism_04.pdf. contains papers read on the 7th Int. Workshop 'Intelligent Agents: Decision-Support and Planning', Udine, Italy, 2004.

D. Geiger, T. Verma, and Pearl J. D-separation: From theorems to algorithms. pages 118–125, Windsor, Ontario, 1989. Association for Uncertainty in Artificial Intelligence, Montain View, CA.

D. Geiger, T. Verma, and Pearl J. Indentifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.

I. J. Good. Explicativity: a mathematical theory of explanation with statistical applications. In *Proceedings of the Royal Societety(London)*, pages 303–330, 1977.

I. J. Good. Weight of evidence: a brief survey. With discussion and a reply by the author. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian statistics*, volume 2, pages 249–269, Valencia, 1985. Elsevier Science Publishers and Valencia University Press.

I. J. Good and W. I. Card. The diagnostic process with special references to error. 1971.

P. Haddawy, J. Jacobson, and C. E. Kahn. Generating explanations and tutorial problems from Bayesian networks. *Journal of the American Medical Informatics Association*, pages 770–774, 1994. Suppl. S.

David Earl Heckerman. *Probabilistic similarity networks*. PhD thesis, Dept of Computer Science and Medicine, Stanford University, 1990.

M. Henrion and M. J. Druzdzel. Qualitative propagation and scenario-based schemes for explaining probabilistic reasoning. In P. B. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 17–32, Amsterdam, 1991. North-Holland.

Max Henrion and Marek J. Druzdzel. Qualitative and linguistic explanation of probabilistic reasoning in belief networks. In *Proceedings of the Third International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 225–227, Paris, France, July 1990.

Eric Horvitz and Matthew Barry. Display of information for time-critical decision making. In Morgan Kaufmann Publishers, editor, *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 296–305, San Francisco, CA, 1995.

Eric Horvitz, John Breese, David Heckerman, David Hovel, and Koos Rommelse. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 256–265, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

R. Howard and J. Matheson. Influence diagrams. In *Readings on Principles and Applications on Decision Analysis*, volume 2, pages 721–762. Strategic Decisions Group, Menlo Park, CA, 1981.

F. V. Jensen, K. G. Olesen, and S. K. Andersent. Bayesian updating in recursive grahical models by local computation. *Networks*, 20:637–659, 1990.

William O Kermack and Anderson G McKendrick. Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. *Proceedings of the Royal society of London. Series A*, 138(834):55–83, 1932.

Jin H. Kim and Judea Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *IJCAI*, pages 190–193, 1983.

Oscar Kipersztok and Haiqin Wang. Another look at sensitivity of Bayesian networks to imprecise probabilities. In *AI and Statistics*, Hyatt Hotel, Key West, Florida, 2001.

D. Koller, U. Lerner, and D. Anguelov. A general algorithm for approximate inference and its application to hybrid Bayes nets. In *Proceedings of the 15th Annual Conference on Uncertainty in Articial Intelligence (UAI-99)*, pages 324–333, San Francisco, CA. Morgan Kaufmann Publishers, 1999.

C. Lacave and J. Diez, F. A review of explanation methods for Bayesian networks. *Knowledge Engineering Review*, 17:107–127, 2002.

C. Lacave, A. Onisko, and F. J. Díez. Debugging medical Bayesian networks with Elvira's explanation capability. In *Workshop on Bayesian Models in Medicine, Eighth European Conference on Artificial Intelligence in Medicine (AIME-2001)*, Cascais, Portugal, 2001.

Carmen Lacave and Francisco Javier Díez. Knowledge acquisition in PROSTANET - A Bayesian network for diagnosing prostate cancer. In Vasile Palade, Robert J. Howlett, and Lakhmi C. Jain, editors, *Knowledge-Based Intelligent Information and Engineering*

Systems, volume 2774 of *Lecture Notes in Computer Science*, pages 1345–1350. Springer, 2003. ISBN 3-540-40804-5.

Carmen Lacave, Roberto Atienza, and Francisco J. Díez. Graphical explanation in Bayesian networks. In *ISMDA'00: Proceedings of the First International Symposium on Medical Data Analysis*, pages 122–129, London, UK, 2000. Springer-Verlag. ISBN 3-540-41089-9.

S. Lauritzen. Propogation of probabilities, means and variances in mixed association models. *Journal of American Statistical Association*, 87(420):1089–1108, 1992.

U. Lerner, E. Segal, and D. Koller. Exact inference in networks with discrete children of continuous parents. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 319–328, San Francisco, CA., 2001. Morgan Kaufmann Publishers.

David Madigan and Krzysztof Mosurski. An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika*, 77(2):315–319, Jun 1990.

David Madigan, Krzysztof Mosurski, and G. Almond, Russell. Graphical explanation in belief networks. *Journal of Computational and Graphical Statistics*, 6(2):160–181, June 1997. ISSN 1061-8600.

Anders L. Madsen, Uffe B. Kjærulff, Jörg Kalwa, Michel Perrier, and Miguel Ángel Sotelo. Applications of probabilistic graphical models to diagnosis and control of autonomous vehicles. *The Second Bayesian Modeling Applications Workshop*, 2004.

J. Mao. *An Experimental Study of the Use and Effect of Hypertext Based Explanation in Knowledge-based Systems*. PhD thesis, University of British Columbia, 1995.

Robert M. Martin. The philosopher's dictionary. Peterborough, Ont. Broadview Press, 1994.

Susan W. McRoy, Alfredo Liu-Perez, Suzan Haller, and James Helwig. B2: A tutoring shell for Bayesian networks that supports natural language interaction. In *Working Notes of the AAAI 96 Spring Symposium on Artificial Intelligence in Medicine*, 1996.

Kathleen Ellen Moffitt. *An empirical test of expert system explanation facility effects on incidental learning and decision-making*. PhD thesis, Arizona State University, 1989. Chairperson-James Hershauer.

I. Nachman, G. Elidan, and N. Friedman. "Ideal parent" structure learning for continuous variable networks. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 400–409, Arlington, VA, 2004. AUAI Press.

E. Neapolitan, Richard. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. John Wiley & Sons, Inc., New York, 1990.

M. Neil, N. E. Fenton, and M. Tailor. Using Bayesian networks to model expected and unexpected operational losses. *Risk Analysis: An International Journal*, 25(4):963–972, 2005.

Norsys Inc. *"Netica v2.17, Software"*. URL www.norsys.com.

Judea Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *AAAI*, pages 133–136, 1982.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 1988.

Nancy Pennington and Reid Hastie. Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 14(3):521–533, July 1988.

Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russell Greiner, David S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. Visual explanation of evidence with additive classifiers. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 21, page 1822. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

K. Wojtek Przytula and Don Thompson. Construction of Bayesian networks for diagnostics. In *Proceedings of 2000 IEEE Aerospace Conference*, 2000.

Silja Renooij and Linda van der Gaag. Enhancing QPNs for trade-off resolution. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 559 – 566, 1999.

Silja Renooij, Linda van der Gaag, Simon Parsons, and Shaw Green. Pivotal pruning of trade-offs in QPNs. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI-00)*, pages 515–5, San Francisco, CA, 2000. Morgan Kaufmann.

K. G. Schulze, R. N. Shelby, D. J. Treacy, and M. C. Wintersgill. Andes: A coached learning environment for classical Newtonian physics. In *Proceedings of the 11th International Conference on College Teaching and Learning*, Jacksonville, FL, April 2000.

E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176, June 2003. ISSN 1061-4036. doi: 10.1038/ng1165. URL http://dx.doi.org/10.1038/ng1165.

Peter Sember and Ingrid Zukerman. Strategies for generating micro explanations for Bayesian belief networks. In *Proceedings of the 5th Workshop on Uncertainty in Artificial Intelligence*, pages 295–302, Windsor, Ontario, 1989.

R. D. Shachter. An ordered examination of influence diagrams. *Management Science*, 35: 535–564, 1990.

A Shafer, G. *Mathematical Theory of Evidence*. Princeton University Press, 1976.

E.H. Shortliffe. *Computer Based Medical Consultations: MYCIN*. American Elsevier, 1976.

M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J.Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: I. the probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30(4):241–255, 1991.

D.J. Spiegelhalter and R.P. Knill-Jones. Statistical and Knowledge-Based Approaches to Clinical Decision-Support Systems, with an Application in Gastroenterology. *Journal of the Royal Statistical Society. Series A (General)*, 147(1):35–77, 1984.

Henri J. Suermondt and Gregory F. Cooper. An evaluation of explanations of probabilistic inference. *Comput. Biomed. Res.*, 26(3):242–254, 1993. ISSN 0010-4809.

Henri Jacques Suermondt. *Explanation in Bayesian belief networks*. PhD thesis, Stanford, CA, USA, 1992.

C.A. Sugar and G.M. James. Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association*, 98(463):750–764, 2003.

William R. Swartout and Johanna D. Moore. Explanation in second generation expert systems. In *Second generation expert systems*, pages 543–585. Springer-Verlag New York, Inc., 1993. ISBN 0-387-56192-7.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001. doi: 10.1111/1467-9868.00293. URL http://www.blackwell-synergy.com/doi/abs/10.1111/1467-9868.00293.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: heuristics and biases. *Science*, 185:1124–1131, 1974.

Michael P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3):257–303, 1990a. ISSN 0004-3702.

Michael P. Wellman. Graphical inference in qualitative probabilistic networks. *Networks*, 20:687–701, 1990b.

Brady West, Kathleen B Welch, and Andrzej T Galecki. *Linear mixed models: a practical guide using statistical software*. CRC Press, 2006.

Wim Wiegerinck. Approximate explanation of reasoning in Bayesian networks. In *Workshop Probabilistic Graphical Models (PGM '04)*, volume 2004, pages 209–216, 2004.

C. Wragg, Edward and George Brown. *Explaining*. Leverhulme Primary Project Classroom skills series. Routledge, Florence, KY, USA, 1993.

Ghim-Eng Yap, Ah-Hwee Tan, and Hwee-Hwa Pang. Explaining inferences in Bayesian networks. *Applied Intelligence*, 29(3):263–278, 2008.

L. Richard Ye and Paul E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, 19(2):157–172, 1995. ISSN 0276-7783. doi: http://dx.doi.org/10.2307/249686.

L. A. Zadeh. Commonsense reasoning based on fuzzy logic. In *WSC '86: Proceedings of the 18th conference on Winter simulation*, pages 445–447, New York, NY, USA, 1986. ACM Press. ISBN 0-911801-11-1.

L. A. Zadeh. Knowledge representation in fuzzy logic. *IEEE Transactions on Knowledge and Data Engineering*, 1(1):89–100, March 1989.

L. A. Zadeh. *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi A. Zadeh.* World Scientific Publishing Co., Inc., 1996.

J. Y. Zhu and A. Deshmukh. Application of Bayesian decision networks to life cycle engineering in Green design and manufacturing. *Engineering Applications of Artificial Intelligence*, 16(2):91–103, March 2003. URL http://www.sciencedirect.com/science/article/B6V2M-495664V-2/2/1d2cd05e69db29a123097021022cb4f1.