Artificial Intelligence in Medicine 53 (2011) 181–204

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Incorporating expert knowledge when learning Bayesian network structure: A medical case study

M. Julia Flores[a,*], Ann E. Nicholson[b], Andrew Brunskill[c], Kevin B. Korb[b], Steven Mascaro[d]

[a] *Departamento de Sistemas Informǐaticos SIMD i^3A, Universidad de Castilla-La Mancha, Campus Universitario s/n, Albacete 02071, Spain*
[b] *Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia*
[c] *University of Washington, Department of Health Services, Box 357660, Seattle, WA 98195-7660, USA*
[d] *Bayesian Intelligence Pty Ltd., 2/21 The Parade, Clarinda, VIC 3169, Australia*

ARTICLE INFO

Article history:
Received 4 March 2010
Received in revised form 28 June 2011
Accepted 9 August 2011

Keywords:
Bayesian networks
Causal discovery
Structure learning
Expert priors
Medical datasets
Heart failure

ABSTRACT

Objectives: Bayesian networks (BNs) are rapidly becoming a leading technology in applied Artificial Intelligence, with many applications in medicine. Both automated learning of BNs and expert elicitation have been used to build these networks, but the potentially more useful combination of these two methods remains underexplored. In this paper we examine a number of approaches to their combination when learning structure and present new techniques for assessing their results.
Methods and materials: Using public-domain medical data, we run an automated causal discovery system, CaMML, which allows the incorporation of multiple kinds of prior expert knowledge into its search, to test and compare unbiased discovery with discovery biased with different kinds of expert opinion. We use adjacency matrices enhanced with numerical and colour labels to assist with the interpretation of the results. We present an algorithm for generating a single BN from a set of learned BNs that incorporates user preferences regarding complexity vs completeness. These techniques are presented as part of the first detailed workflow for hybrid structure learning within the broader knowledge engineering process.
Results: The detailed knowledge engineering workflow is shown to be useful for structuring a complex iterative BN development process. The adjacency matrices make it clear that for our medical case study using the IOWA dataset, the simplest kind of prior information (partially sorting variables into tiers) was more effective in aiding model discovery than either using no prior information or using more sophisticated and detailed expert priors. The method for generating a single BN captures relationships that would be overlooked by other approaches in the literature.
Conclusion: Hybrid causal learning of BNs is an important emerging technology. We present methods for incorporating it into the knowledge engineering process, including visualisation and analysis of the learned networks.

1. Introduction

Bayesian networks (BNs) [1,2] are rapidly becoming a leading technology in applied Artificial Intelligence. By combining a graphical representation of the dependencies between variables with probability theory and efficient inference algorithms, BNs provide a powerful and flexible tool for reasoning under uncertainty. Medicine is a complex domain where experienced medical practitioners hold much of their knowledge implicitly, making an appealing target for expert systems development. BNs are particularly suited to medical applications due to their ability to explicitly model causal interventions, to reason both diagnostically and predictively, and to help visualise their relations graphically, assisting with their understanding.

BNs may be built either by eliciting expert knowledge or by automated causal discovery. There have been many medical applications of BNs, including early elicited networks such as the ALARM network for monitoring patients in intensive care [3], diagnosing oesophageal cancer [4], mammography [5] and diagnosing liver disorder [6]; see [2, Ch. 5] for a recent survey. To avoid the long development times required by expert elicitation, other

[*] Corresponding author. Tel.: +34 967599284; fax: +34 967599224.
E-mail addresses: Julia.Flores@uclm.es (M. Julia Flores), Ann.Nicholson@monash.edu (A.E. Nicholson).

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.08.004

approaches have been taken. In previous research we built BNs for predicting coronary heart disease in two other ways [7,8]: (1) knowledge engineering BNs from the medical literature, using published epidemiological models of coronary heart disease, supplemented by medical expertise to clarify interpretation; and (2) applying the causal discovery program CaMML [9,10] to learn BNs from data from the Australian Busselton study [11]. Other examples of medical BNs learnt from data include emergency medical service data [12] and tuberculosis epidemiology [13]. The PROMEDAS medical decision support tool [14] also uses BNs, automatically compiling both the network and an interface from the underlying medical database (which currently covers the areas of endocrinology and lymphoma diagnostics).

Both approaches to building BNs have limitations: expert elicitation is expensive, time-consuming and relies on experts having full domain knowledge, while automated learning is often ineffective given small or noisy datasets. This has led to hybrid approaches which incorporate prior expert information into the causal discovery process. For example, the Tetrad IV BN learner can use information about causal "tiers" [15], while CaMML has been extended to use more diverse forms of expert information [16]. One application of such a hybrid approach, using information obtained from textual analysis of the literature together with learning from data, has produced a BN clinical model of ovarian cancer [17]. However, to our knowledge, different kinds of expert elicited structural priors to inform causal discovery have not been used in medical or other applications to date. One reason is that most BN learning software provides only limited support for it. Another is that while methodologies have been developed for the knowledge engineering of BNs (e.g., [18,19]), no detailed methodology is available for building the network structure using the hybrid approach.

In this paper we present a detailed case study of the use of expert priors in combination with automated causal learning to model heart failure, using the so-called Iowa dataset [20]. Heart failure is a difficult and complex problem, and this publicly available dataset has many limitations, including no variable for the actual heart failure diagnosis, and no recording of times for past events. Thus, the major contribution of this paper is not new results for the domain, but instead the presentation of a comprehensive methodology for this hybrid approach to constructing BNs, together with new ways of analysing and visualising these structures.

In Section 2 we describe the medical case study, including the description of the dataset and preliminary data analysis to select attributes/variables, and give some background on "knowledge engineering" of BNs. In Section 3 we describe CaMML and how it can incorporate different types of expert knowledge. We then describe all the steps in the knowledge engineering process we undertook for the case study, including data preprocessing (Section 4.1), the domain information elicited from our expert (Sections 4.2–4.4) and our experimental methods for incorporating those priors into CaMML (Section 5). In Section 6 we systematically test and compare unbiased discovery with discovery biased with different kinds of expert opinion; the results and their analysis are presented using graphs, edit distance and new kinds of adjacency matrices for enhancing comparisons. In Section 5.7 we demonstrate that a BN scoring metric can be used to generate a single BN from a set of learned BNs, trading off complexity vs completeness.

2. Background

2.1. Case study

Our case study involves a public dataset, the so-called Iowa dataset [20]. Last updated in 1998, this dataset contains 14,456 records and reports data from patients in four different US states. The Inter-University Consortium for Political and Social Research (ICPSR) from the University of Michigan published their research results based on distinct questionnaires answered between 1981 and 1993. The id number of the ICPRS study is 9915; see http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/09915.xml[1] for more detailed information. From this study we have focused on part I, called *Baseline Data*. The full IOWA dataset contains 253 variables, spanning an enormous range of medical conditions. We decided to limit the case study to a single condition, heart failure, chosen because it is common, costly, disabling and deadly.

Heart failure is a disorder that may result from any structural or functional disorder that decreases the ability of the heart ventricle to fill with or eject blood. Although there are numerous ways of assessing cardiac function there is no single diagnostic test, which makes diagnosis difficult, and at least six scoring methodologies have been developed to assess it. Investigations using echocardiography have found that only 50% of participants with left ventricular dysfunction are symptomatic, hence the diagnosis is largely based on a careful history and physical examination [21]. The value of different symptoms in predicting heart failure in a study of referred patients [22] had suggested that while dyspnoea on exercise is a sensitive measure of heart failure (100%) it is not specific (17%) while a history of paroxysmal nocturnal dyspnoea (which is severe shortness of breath occurring at night when the patient is lying down) is only 39% sensitive but has a markedly higher specificity (80%).

Essentially heart failure is a probabilistic clinical diagnosis relating to a constellation of symptoms occurring in a variety of conditions stressing the heart with no definitive objective test. Thus, unsurprisingly, the Iowa dataset does not contain a variable for heart failure diagnosis; instead we model the relationships between the relevant variables representing background factors, past health and current symptoms.

In developing countries around 2% of adults suffer from heart failure, with this figure increasing to 6–10% for those over the age of 65 [23]. Heart failure is associated with high health expenditure, especially the costs of hospitalisation. It is also associated with significantly reduced physical and mental health, resulting in a markedly decreased quality of life. The introduction of learned models with the ability to incorporate expert priors offers the possibility of extending our understanding of the factors contributing to heart failure in community settings.

2.2. Bayesian networks

A Bayesian network [1] is a directed acyclic graph (DAG) whose nodes represent random variables and arcs represent direct dependencies (e.g., causal relationships). Each node has a conditional probability table, quantifying the relationship between connected variables. Users can set the values of any combination of nodes in the network that they have observed. This evidence propagates through the network, producing a new probability distribution over all the variables in the network. There are a number of efficient exact and approximate inference algorithms for performing this probabilistic updating, providing a powerful combination of predictive, diagnostic and explanatory reasoning. Fig. 1(a) shows an example BN, a variant of the well-known "Asia" BN from [24].

2.2.1. BNs for medical applications

The simplest structure for a medical diagnostic BN is the so-called naive Bayes model, shown in Fig. 1(b), where the

[1] Last access on 20 January 2010.

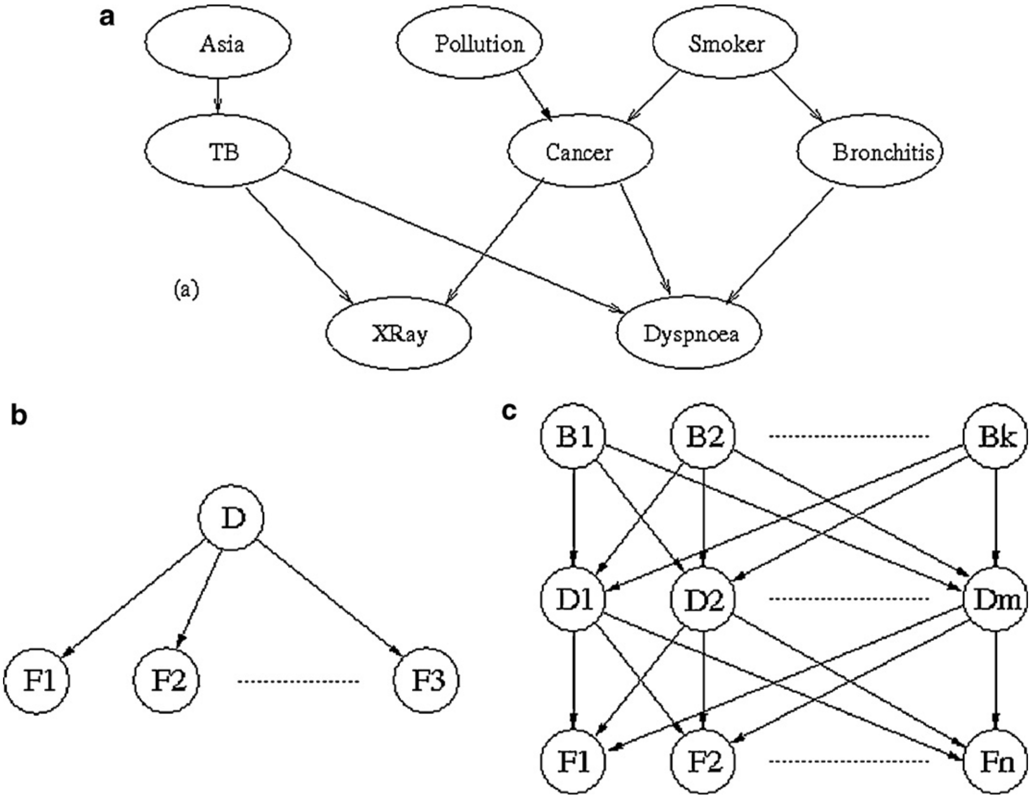


Fig. 1. (a) Example BN for the "Asia network", for the scenario: A patient has been suffering from shortness of breath (called dyspnoea) and visits the doctor, worried that he has lung cancer. The doctor knows that other diseases, such as tuberculosis (TB) and bronchitis, are also possible causes. She also wants to find out whether or not the patient is a smoker (increasing the chances of cancer and bronchitis) or has recently visited Asia (as TB is more prevalent there). A positive X-ray would indicate either TB or lung cancer. (b) Generic BN structures for medical diagnosis, naive Bayes model; (c) multiply connected network [2, Fig. 5.1].

Disease (D) node has values for the candidate diseases, while *Findings (F)* nodes represent both symptoms and test results. There are two simplifying assumptions in this network that often go wrong: that the patient can have only a single disease and that symptoms are independent of each other given the disease.

The multiply connected network of Fig. 1(c) is a more realistic, but clearly more complex, model. It contains a Boolean node for each disease under consideration, while the *Background (B)* nodes represent patient history, such as the age, sex and smoking. While avoiding the distorting simplifications above, in practice this structure is likely to be too complex, requiring probabilities for the combined effect of every disease on each finding. An example of this two-level network structure is that of the "quick medical response decision-theoretic (QMR-DT)" project [25], a probabilistic version of the frame-based CPCS knowledge base for internal medicine. In QMR-DT, the problem of complexity was ameliorated with the so-called binary 'noisy-or' model, where it is assumed the effect of a disease on its findings is independent of other diseases and findings. One version of the QMR-DT network (described in [26]) had 448 nodes and 908 arcs, including 74 background nodes (which they called "predisposing factors"); more than 600 probabilities were estimated, a large but not an unreasonable number given the scope of the application.

In this paper, we are looking at building BNs for a specific medical condition (e.g., heart failure), where we have a fair amount of data, from which we will learn both the structures and the parameters. However the crucial insight is that in a properly structured causal BN, the background factors should be parents (or ancestors) of the diseases, which in turn are ancestors of the physical effects they generate.

2.2.2. Knowledge Engineering Bayesian networks (KEBN)

Knowledge Engineering can be viewed as an engineering discipline that involves integrating knowledge into computer systems in order to solve problems normally requiring a high level of human expertise. The aim is to construct a so-called 'expert system', a model able to give answers similar to those of an expert human. In our case study, the result of the knowledge engineering process will be an expert system consisting of a Bayesian network (with the knowledge represented in the graph structure and the conditional probabilities) and an ability to reason given evidence, using powerful belief updating techniques (e.g., [27]).

When developing a BN for a specific domain, the knowledge engineering process involves constructing a model that is sufficiently complex to realistically represent the problem features, while being simple enough to be tractable in terms of both development and application. The KEBN life cycle [18, Fig. 6.1], as reproduced in Fig. 2, describes at a high level the initial development of a network, utilising any number of iterations over its stages, each iteration producing some refinement of the network. KEBN methods include both elicitation from experts and learning from data; the combination is the focus of this paper. The development of any BN will follow a particular path through this generic workflow; in Section 5.1 we discuss the detailed KEBN process for Stage 1 we followed.

2.2.3. Learning Bayesian networks

There is one key limitation when learning BNs from observational data only—there is usually no unique BN that represents the joint distribution. More formally, two BNs in the same *statistical equivalence class* (SEC) [28] can be parameterised to give an identical joint probability distribution. There is no way to distinguish

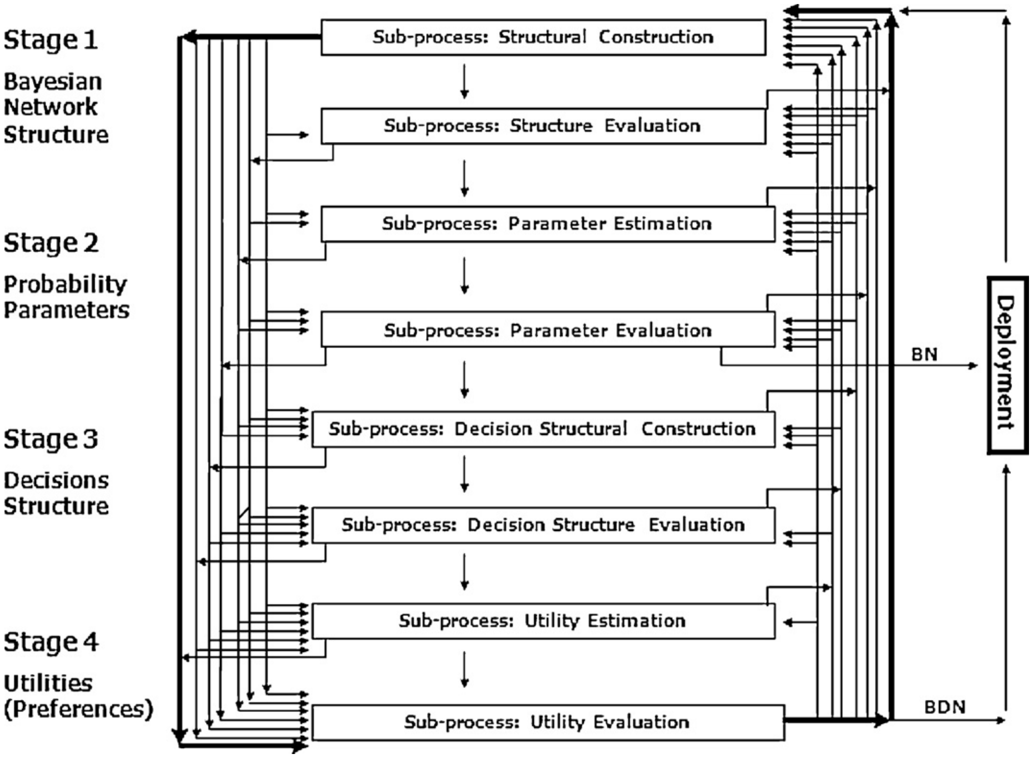

Stage 1
Bayesian Network Structure
Stage 2
Probability Parameters
Stage 3
Decisions Structure
Stage 4
Utilities (Preferences)
Sub-process: Structural Construction
Sub-process: Structure Evaluation
Sub-process: Parameter Estimation
Sub-process: Parameter Evaluation
Sub-process: Decision Structural Construction
Sub-process: Decision Structure Evaluation
Sub-process: Utility Estimation
Sub-process: Utility Evaluation
BN
BDN
Deployment

Fig. 2. KEBN life cycle [18, fig. 6.1].

between the two using only observational data (although they may be distinguished given experimental data).

BN structural learning algorithms can be classified into constraint-based and metric-based. Constraint-based methods (e.g., PC [29], RAI [30]) use information about conditional independencies gained by performing statistical significance tests on the data. Metric-based methods (e.g., K2 [31], CaMML [32]) search for a BN to minimise or maximise a metric; many different metrics have been used, (e.g., K2 uses the BDe metric, CaMML uses an MML metric; see [2, Ch 9]). Metric-based BN structural learners also vary in the search method used and in what is returned from the search; some learners (e.g., K2) return a DAG, others (e.g., GES [33]) learn only the SEC. The metric-based methods can incorporate expert knowledge about the relationships between variables – the focus of this paper – by using them as structural priors that alter the "score" given to a BN. Here we use CaMML, as it provides more types of structural priors (described in the next section) than any other BN learner, metric or constraint based.

These methods have been applied to learning medical BN applications. Acid et al. [12] compare 4 different BN learning algorithms (including PC and two metric based methods), using emergency medical service data. Antal et al. [17] use a Bayesian metric learning method combined with information from the literature to build BNs as clinical models of ovarian tumours. Getoor et al. use a metric-based method for learning both BNs, and a variant called "statistical relational models", using epidemiological data on tuberculosis, while CaMML has been applied [7] to the Australian Busselton data [11].

3. CaMML: a tool for learning BNs

CaMML attempts to learn the best causal structure to account for the data, using a minimum message length (MML) metric with a two-phase search, simulated annealing followed by Markov Chain Monte Carlo (MCMC) search, over the model space. Both MML [34,35] and the better known MDL [36] are inspired by information theory, and make a tradeoff between prior probability (model complexity) and goodness of fit. With both, the problem becomes one of encoding both the model and the data, and the best model is then one that minimises the message length for that encoding. The differences between MDL and MML are largely ideological: MDL is offered specifically as a non-Bayesian inference method, which eschews the probabilistic interpretation of its code, whereas MML specifically is a Bayesian technique.[2] Across a range of problems, CaMML has matched the best alternative programs [39–41].

The full details of MML encoding are not required for this paper, but we can write the relationship between the message length, the model and the data given the model as:

$$msgLen \propto -\log(P(Model)) - \log(P(Data|Model)).$$

The CaMML metric is a combination of the message of the MML encoding of the BN, incorporating three parts: (1) the network structure, (2) the parameters given this structure, and (3) the data given the network structure and these parameters.

O'Donnell et al. [16,42] modified the original CaMML encoding of the BN structure to incorporate expert priors. We now look at the forms of these structural priors.

CaMML supports multiple ways of describing prior information about relationships between variables. While both specific and accurate information is ideal, we generally prefer accuracy over specificity. For example, an expert might know that there is a dependency between chronic heart disease and smoking, without

[2] Lam and Bacchus [37] developed the first MDL encoding for BNs, while Suzuki [38] proposed an alternative MDL implementation (see also [2, Ch 9.3]); neither is available in a BN learner software package.

knowing through which variables it operates. If the information we require is too specific, an expert may refuse to give anything useful. Hence, CaMML allows several levels of structural information, each of which can be accompanied by a confidence level. These levels, from most specific to most general, are:

Full structure. An expert may supply a fully specified network. Alternatively, an expert may supply a proper subnetwork, describing only the variable relations he or she is confident about.

Direct causal connections between variables may be indicated (e.g., $A \rightarrow B$). This requires a high level of knowledge in the workings of the causal processes between the variables.

Direct relation (A–B). It may be known that two variables are related directly, but the direction of causality is unknown.

Causal dependency ($A \Rightarrow B$). This allows an expert to indicate that one variable is an ancestor of the other, when the mechanism between them remains unknown. For example, it is generally accepted that smoking causes lung cancer, however little is known about the detailed process.

Temporal order (tiers) ($A \prec B$). In many domains it is clear that some variables come before others; CaMML allows that to be indicated independent of other information. More generally, CaMML allows us to specify causal *tiers*—that is, causal relationships between sets of variables, based on the notion that tiers separate the variables on a timeline and that causality only occurs in the forward time direction. This use of "temporal tiers" to give a partial ordering that constrains arcs in a causal network is standard in BN learning (e.g., Tetrad IV and K2). For example: $\{A_1, A_2, ...A_m\} \prec \{B_1, B_2, ...B_n\}$ where $\prec$ means that the variables A_i occur before the B_j, which for CaMML becomes the structural constraint the A_i cannot be descendants of the variables B_j. Note that this is weaker than defining a tier in terms of ancestors, since there is no need to have directed paths from each A_i to each B_j (or, indeed, any paths at all between A's and B's).

Correlation ($A \sim B$). The most general sort of information CaMML uses is correlation. This implies that there is some connection between the nodes. It may be a causal dependency in either direction or via a common ancestor.

Other BN learners also support the use of structural knowledge, but they have been limited to specifying one or two of these kinds of priors. K2 [31] *requires* a total ordering; Heckerman and Geiger [43] relax this requirement by proposing the use of a Minimum Weighted Spanning Tree (MWST) algorithm to learn a tree-like BN structure, which can then be used to initialise K2. Tetrad IV and GeNIe both support tiers, Hugin Lite [44] allows the user to provide directed arcs, while BDe/BGe require full or partial networks to be specified [45]. Antal et al. [17] use only undirected pairwise priors, comparing "text-based" priors obtained by statistical analysis of the literature to expert elicited priors. However, these are all hard constraints. Heckerman and Geiger [43] proposed soft constraints by computing a prior distribution from the edit distance between an expert specified structure and the candidate structure, while Castelo and Siebes [46] use a prior over directed arcs. To our knowledge, neither soft constraint method is yet supported in any BN structural learning package.

With CaMML the expert may provide a full structure or any combination of the prior types above. Experts may also provide their confidence for each piece of information they are providing. While this confidence is provided to CaMML in the form of a probability, other forms may be used during elicitation (e.g., a discrete numerical scale, or a verbal scale) and later mapped into probabilities. CaMML then uses a system based on the MML encoding of each kind of structural prior and the confidence probabilities, together with the default arc probability, to synthesise the information into a coherent whole.

4. Structural priors for the case study

4.1. Pre-processing

From the original Iowa dataset, our medical expert[3] selected those variables he considered most relevant for the present study. As noted earlier, there is no single variable representing a diagnosis of heart failure. The selected variables span a combination of patient background, relevant medical history and current symptoms. For example, SHBRLIE represents severe shortness of breath occurring at night when the patient is lying down, which is likely to be most specifically a symptom of congestive heart failure albeit occurring only in a minority of patients. The 253 variables in the original dataset were whittled down to 28, with all 14,456 cases being retained. The raw data was then converted to Weka's ARFF format, which is also used by CaMML. Table 1 shows descriptions and possible values for the chosen variables. (The last column indicates the tier to which each variable was initially assigned; these tiers will be explained in Section 4.2). We can see from this table that the dataset holds only patient reported data and does not provide times of past events, however this is a common problem with medical datasets.

Pre-processing is a crucial step in data mining [47, Ch. 2]; it is clear that without useful data, there can be no useful results! Among the many pre-processing tasks that can be performed, we undertook the following four, which were required for our purposes.

4.1.1. Data cleaning

There were a number of missing attribute values in the data set. It proved worthwhile working directly with the expert because many of these missing values were in fact known. In many such cases, a missing value indicated a negative answer, in particular for the cases: HADHRTAT (56 missing values), ABLWALK (469), HADSTRKE (29), HADCANCR (23), HADDBTES (30), TKINSLN (90% of the values for this variable were missing and the expert suggested that this clearly indicated a negative answer, plus where there were values, less than 5% were positive) HADHGHBL (41), TKMEDBLD (50%; the expert's reasoning was the same as for TKINSLN), PAINWLK (33%), TKMEBLD (55%) and SHRBRLIE (6%).

Following the recommendation of our expert, missing entries for the variable EVRSMKCG were supplied with the new state *Unknown*. Finally, for the remaining variables we used imputation techniques to provide the missing value. These were WEIGHT (889), HEIGHT (1324), SCDSYSTL (1232), SCDDSTLC (1238) and SMKCGNW (59).

4.1.2. Normalization and aggregation

Examining the previous variables, it does not seem reasonable to treat WEIGHT and HEIGHT as distinct factors. Separately, they do not allow us to say very much about the possibility of heart failure; combined, they are much more valuable. Conveniently, there is a already an index for combining weight and height that is commonly accepted in the medical world: the Body Mass Index or BMI. Given that the dataset is based on US units of pounds and inches, BMI was computed by $WEIGHT \times 703/HEIGHT^2$ and aggregated according to the categories given in http://www.nhlbisupport.com/bmi/, that is: Underweight, if $BMI \leq 18.5$; Normal, if $BMI \in (18.5,25)$; Overweight, if $BMI \in [25,30)$; and Obesity, if $BMI \geq 30$. In the dataset, we call this new attribute the *BMIKind*.

[3] One of the authors, Dr Andrew Brunskill.

Table 1
Description of the attributes and assignment to tiers.

Nr	Name	Description	Values	Tier
i	SITE	Location	1 = East Boston, 2 = Iowa,	bg2
ii	ID	Subject ID	3 = New Haven, 4 = Duke numeric	
iii	SEX	Gender	1 = male, 2 = female	bg1
iv	RACE	Race	1 = White & other Non-black, 2 = black	bg2
v	AGEINT	Age of the subject	1 = age <70, 2 = 70–74, 3 = 75–79, 4 = 80–84, 5 = age > 85	bg1
vi	EVRMARRD	*Have you ever been married?*	1 = Yes, 2 = No	bg2
vii	MARTSTAT	Marital status	1 = NowMarried,2 = Separated,3 = Divorced	bg2
viii	CURRWRK	*Currently working at paying job*	1 = Yes, 2 = No	ci
ix	RETIRED	*Are you retired?*	1 = Yes, 2 = No	ci
x	AGE	Answer to *How old are you?*	1 = Incorrect, 2 = Correct, 3 = Refused	–
xi	NDHLPWLK	*Need help—walk in doors?*	1 = NoHelp, 2 = Help, 3 = Unable	ci
xii	ABLWALK	*Able to walk 1/2 mile without help*	1 = Yes, 2 = No	ci
xiii	HADHRTAT	*Dr ever told you had heart attack?*	1 = Yes, 2 = Suspect, 3 = No	p
xiv	HADSTRKE	*Dr ever told you had stroke?*	1 = Yes, 2 = Suspect, 3 = No	p
xv	HADCANCR	*Dr ever told you had cancer?*	1 = Yes, 2 = Suspect, 3 = No	p
xvi	TKINSLN	*Currently taking insulin*	1 = Yes, 2 = No	cd
xvii	HADDBTES	*Dr ever told you had diabetes?*	1 = Yes, 2 = Suspect, 3 = No	p
xviii	HADHGHBL	*Dr ever told you high blood pressure?*	1 = Yes, 2 = Suspect, 3 = No	p
xix	TKMEDBLD	*Currently taking medication for high blood*	1 = Yes, 2 = No	cd
xx	PAINWLK	*Get pain when walking at ordinary pace?*	1 = Yes, 2 = No	cd
xxi	PRSSCHST	*Ever had any pressure in chest?*	1 = Yes, 2 = No	cd
xxii	WEIGHT	Weight in pounds	Numeric	–
xxiii	HEIGHT	Height in inches	Numeric	–
xxiv	SCDSYSTL	*Second blood pressure reading-Systolic*	Numeric	cd
xxv	SCDDSTLC	*Second blood pressure reading-Diastolic*	Numeric	cd
xxvi	SMKCGNW	*Do you smoke cigarettes (regularly) now?*	1 = Yes, 2 = No	–
xxvii	EVRSMKCG	*Did you ever smoke cigarettes (regularly)?*	1 = Yes, 2 = No	bg2
xxviii	SHRBRLIE	*Short of breath when you are lying flat?*	1 = Yes, 2 = No	cd

4.1.3. Data reduction

The number of variables in the original dataset was already reduced by our expert to 28. However, there was a further attribute that was clearly unnecessary for our task: the person identification number ID. This would be unique in each instance, and hence provide no predictive value.

4.1.4. Numerical data discretisation

We did not directly perform this task in the preprocessing phase, since many of the variables were already discrete. The discretisation of BMI into the attribute BMIKind was described earlier. The only remaining numeric variables were those related to blood pressure. We left those as they were since CaMML can discretise numerical attributes directly, using the method most suited to the data (supervised or unsupervised).

By performing the above pre-processing tasks, we produced a dataset, with no missing values, that was suitable for use in the learning experiments. Here, we call this dataset *DataHF*.

4.2. Grouping into tiers

In Section 2.2.1 we saw a generic BN causal structure for medical applications: background factors are parents of disease nodes which are parents of finding nodes. However, this structure is too simplistic for most real cases; for example, there are often temporal relationships or complex interactions between background factors, between diseases, or between symptoms and test results.

However, as we have seen, CaMML can capture some expert domain knowledge in the form of partial orderings based on the notion of 'temporal tiers', such as groups for background, disease and findings. We now describe how we applied this approach to the case study.

4.2.1. Background variables

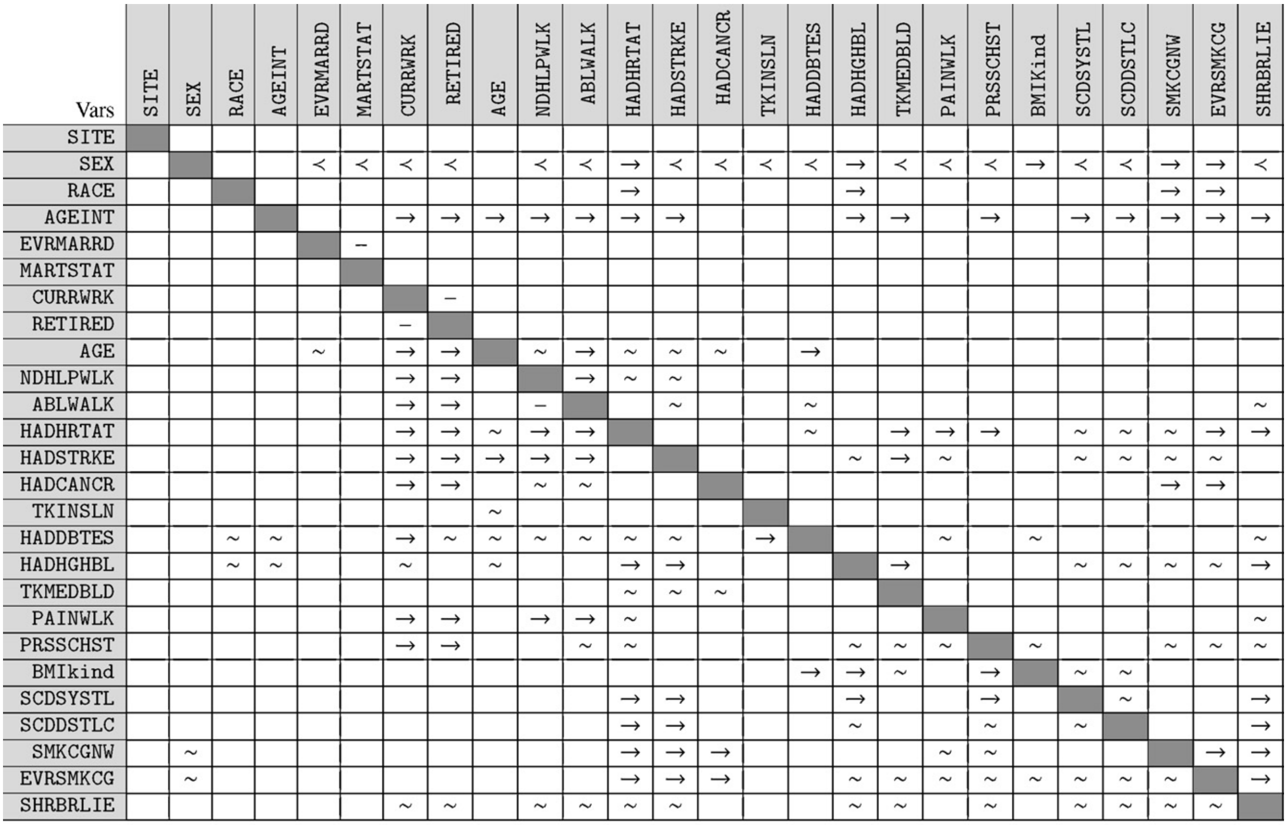
The background variables for this dataset are: location (i), sex (iii), race (iv), age (v), ever married (vi), marital status (vii) and ever smoked regularly (xxvii). We further postulated that of these age (v) and sex (iii) are more fundamental, potentially influencing (though not directly causing) the other background factors. Therefore, in Table 1, we label background variables as type bg, with bg1 referring to the more fundamental background variables and bg2 referring to the remainder.

4.2.2. Disease and finding variables

In our case study, it was less clear how to choose meaningful disease variables. This dataset contains information about what doctors have told the patient in the past, including events (heart attack, stroke), disease (cancer, diabetes) and symptoms (high blood pressure), rather than direct information about a patient's current disease. Therefore, we decided to divide non-background variables into 'past' and 'current'. The 'past' variables are: had heart attack (xiii), had stroke (xiv), had cancer (xv), had diabetes (xvii) and had high blood pressure (xviii). These are shown as type p in Table 1. 'Current' variables, which correspond to "findings" in the generic medical BN (see Fig. 1(c) were further divided into direct and indirect indicators of current health. Direct symptoms are: currently taking insulin (xvi), currently taking medication for high blood pressure (xix), pain walking (xx), ever had pressure in chest (xxi), shortness of breath when lying flat (xxviii), second blood pressure readings (xxiv, xxv). Indirect indicators are: currently working (viii), whether retired (ix), need help to walk indoors (xi), and able to walk 1/2 mile without help (xii). These current health indicators are shown as type c in Table 1, with cd for direct indicators and ci for the indirect indicators.

4.2.3. Unclassified variables

There were some variables that did not fall neatly into any of the above categories. These non-tiered variables are marked with "–" in Table 1. Note that weight (xxii) and height (xxiii) were removed from the dataset, while BMIKind was not added to any tier.



Vars	SITE	SEX	RACE	AGEINT	EVRMARRD	MARTSTAT	CURRWRK	RETIRED	AGE	NDHLPWLK	ABLWALK	HADHRTAT	HADSTRKE	HADCANCR	TKINSLN	HADDBTES	HADHGHBL	TKMEDBLD	PAINWLK	PRSSCHST	BMIkind	SCDSYSTL	SCDDSTLC	SMKCGNW	EVRSMKCG	SHRBRLIE
SITE	■																									
SEX		■			≺	≺	≺	≺		≺	≺	→	≺	≺	≺	≺	→	≺	≺	≺	→	≺	≺	→	→	≺
RACE			■									→					→							→	→	
AGEINT				■			→	→	→	→	→	→	→				→	→		→		→	→	→	→	→
EVRMARRD					■	–																				
MARTSTAT						■																				
CURRWRK							■	–																		
RETIRED							–	■																		
AGE					~		→	→	■	~	→	~	~	~		→										
NDHLPWLK							→	→		■	→	~	~													
ABLWALK							→	→		–	■		~			~										~
HADHRTAT							→	→	~	→	→	■				~		→	→	→		~	~	~	→	→
HADSTRKE							→	→	→	→	→		■				~	→	~			~	~	~	~	
HADCANCR							→	→		~	~			■										→	→	
TKINSLN									~						■											
HADDBTES			~	~			→	~	~	~	~	~	~		→	■			~		~					~
HADHGHBL			~	~			~		~			→	→				■	→				~	~	~	~	→
TKMEDBLD												~	~	~				■								
PAINWLK							→	→		→	→	~							■							~
PRSSCHST							→	→			~	~					~	~	~	■	~			~	~	~
BMIkind																→	→	~		→	■	~	~			
SCDSYSTL												→	→				→			→		■	~			→
SCDDSTLC												→	→				~			~		~	■			→
SMKCGNW		~										→	→	→					~	~				■	→	→
EVRSMKCG		~										→	→	→			~	~	~	~	~	~	~	~	■	→
SHRBRLIE							~	~		~	~	~	~				~	~		~		~	~	~	~	■

Key: *R* is the row variable, *C* is the column variable.
→ *R* directly causes *C*
≺ *R* occurs before *C* (a partial ordering, *R* cannot be a descendent of *C*)
– *R* and *C* are directly related (one causes the other, but order not known)
∼ *R* and *C* are correlated.

Fig. 3. Pairwise relations obtained through the expert elicitation process.

4.2.4. *Tiers*

Given this classification of variables into sets bg = bg1∪bg2, p, and c = cd∪ci, the natural tiers that arise are:

$$bg1 \prec bg2 \tag{1}$$

$$bg \prec p \prec c \tag{2}$$

4.3. *Expert priors on pairwise relationships*

In this study we focused on capturing expert understanding of relationships between variables, and did not attempt to capture a full or partial BN structure from the expert. Instead, we used the following pairwise elicitation method. The expert was provided with a cross-table of 26 rows and 26 columns. For each cell in row *R* and column *C*, the expert indicated the relation, if any, that he believed exists between variable *R* and variable *C*. We briefly described to the expert the difference between the types of structural information CaMML takes: →, –, ⇒, ≺ and ∼. For the relations where order matters (→, ⇒, ≺) the direction was always *Row* relationship *Col*. Note that we could have given the expert only the upper triangle, adding complexity with additional symbols for the reverse direction relationship, although reducing the possibilities of inconsistent information. In order to reduce the elicitation burden on the expert, we did not elicit a confidence in the relationship; we investigated this parameter experimentally. We also note that the expert was located remotely from the knowledge engineers and most communications were electronic.

The result of the expert elicitation was the set of priors for pairwise relations shown in Fig. 3. We expected that most cells would indicate no relation, and this was indeed the case, with 444 empty cells out of 676 (65.68%). Note also that when our expert indicated a non-directional relationship (i.e., – or ∼), say in cell (*R*,*C*), we did not require him to duplicate it in (*C*,*R*) (although he did so in a few cases). In summary, there were 84 →, 4 – (plus one symmetrical not explicitly indicated), 16 ≺ and 89 ∼ (plus the 38 symmetrical ones not explicitly indicated) pairwise relationships specified by the expert, which provided 232 of the 676 possible pairwise relationships. Note that there are no ⇒ relationships, as our expert did not consider non-direct causality a natural relationship to specify. The elicitation was done iteratively; that is, the initial table provided by the expert was checked by the knowledge engineers with inconsistencies identified and some entries queried because they seemed to contradict "commonsense". Interestingly, the elicitation process highlighted differences between medical and causal BN terminology. For example, in causal BN terminology, when a variable is said to directly precede another, that means that in a causal process the first one comes before the other. However, we found that our expert initially confused this with the diagnostic reasoning process, with the effect "leading to" the cause. For example, the expert used TKINSLN → HADDBTES, to represent "if you are taking insulin this means that you had diabetes". However, in causal language having diabetes results in taking insulin, represented by HADDBTES → TKINSLN. This confusion as to whether a BN arc represents the causal process or the diagnostic reasoning has been seen in other BN modelling case studies (e.g., [48]). The expert revised the pairwise relations after the clarification of the relationship semantics, and as a result of discussions with the knowledge engineers.

We found that there were no outright contradictions in the expert's pair-wise priors, however in some cases the expert suggested both $A \rightarrow B$ and $B \sim A$, which CaMML does not allow. We must choose one or the other, since the second relation is more general than the first (and thus includes it). When this occurred, we chose to keep the more restrictive relation, $A \rightarrow B$.

The expert also specified two cases that appeared unusual to the knowledge engineers: HADHRTAT → EVRSMKCG and AGEINT → AGE. When raised with the expert, he indicated that these relationships were misunderstandings. For HADHRTAT → EVRSMKCG, the expert suggested that *it is quite possible to have a heart attack without a history of smoking and to have a history of smoking without having had a heart attack*. He suspected that he initially included this relationship because of clear evidence showing that suffering a heart attack is a potent reason for giving up smoking (i.e., for someone to move from SMKCGNW to EVRSMKCG) but this is a separate issue. For AGEINT → AGE, the expert agreed that the relationship was erroneous and he had been confused when it was introduced. He added that the only issue that might have given rise to this response was that AGEINT is a partial check for whether the patient has correctly recalled their age. If a patient cannot recall their age correctly, this suggests that their other answers may be unreliable or incorrect or they may be having early memory or dementia problems (or both). However, these are not sufficient grounds for establishing a direct causal edge in this case.

Ordinarily, these discoveries would suggest repeating the experiment after removing the two incorrect priors. We decided not to do so, in part to keep the volume of results manageable, but also because we wanted to see whether combining the tier and the pair-wise priors is a way to overcome incorrect priors.

Another common way to validate expert opinion is to elicit information independently (using the same elicitation method) from multiple experts, and use a process such as the Delphi method [49] to reach a consensus. Our decision to use a single expert was dictated partly by practical considerations – we had only a single expert at our disposal – and partly because the focus of this study was on combining different types of expert priors with data in the BN learning process, and we wanted to avoid adding the additional complexity to the KEBN process.

4.4. Combined priors

Intuitively, it seems reasonable to use a combination of tier priors and expert pairwise priors, rather than either set of priors alone. Our early experimental results backed this intuition. Several interesting issues arose when we merged these two sources of prior information, discussed below.

4.4.1. Redundant priors

Many of the priors indicated by the expert were already present in the tiers; for example, SEX precedes all the rest in the tier prior, making redundant the expert's 21 $\prec$ and → pair-wise priors.

4.4.2. Tier violations

There were three relationships provided by the expert that violated tier information, namely, SCDSYSTL → HADHRTAT, SCDSYSTL → HADSTRKE and SCDSYSTL → HADHGHBL. Note that these might be further examples of the expert specifying arcs in the *diagnostic* direction; e.g., in answer to the question, "What does the SCDSYSTL reading tell us about the patient's past?" But it may also be that high blood pressure is a chronic condition, therefore current high readings are very likely to indicate high pressure also in the past, which is clearly causally related to the occurrence of the cardiovascular events. This demonstrates how complex the assignment of causal relationships with undated non-specific and gradual onset symptoms can be. In this case, we decided to use the tiers priors for the combined prior experiments.

5. Experimental methodology

5.1. Our application of the KEBN process

First we describe how we applied the KEBN process (Section 2.2.2) in our case study. Our workflow is pictured in Fig. 4, showing how we are focusing on finding the BN *structure* (Stage 1); other stages of the KEBN process (including parametrisation) were omitted.

The input to our process is simply the IOWA dataset presented in Section 2.1. First, we preprocessed the data by cleaning it up, removing inconsistencies, reducing the number of variables, defining a new summary variable (BMIKind) and discretising continuous variables (Section 4.1); each of these is depicted as an alternative method during preprocessing (and, of course, others could be added).

The preprocessed data is then given to CaMML, our BN learner; the KEBN workflow diagram shows the alternatives available of using data only, or first eliciting either tier priors or pair-wise priors. Priors are given to provided to CaMML together with confidences.

CaMML has a number of parameters that can be adjusted, including the number of runs and how many BNs to be returned from each run. This is represented in the "Set Learner Config" step.

Finally, once CaMML has been run over the data with the appropriate configuration for each experiment, it outputs the learned BNs. We analysed the BN *structure* using the range of measures and visualisations described in Sections 5.4 and 5.6, then generated a single BN from the set of learned BNs (Section 5.7).

Note that KEBN is fundamentally iterative, and our experience in this study was no different. Fig. 4 (below) shows the paths taken through the workflow diagram for each iteration. These paths roughly correspond to the experiments described in the following section; the order of these paths and the choices for each iteration were motivated by analysis of previous results.

5.2. CaMML configuration

CaMML has a number of parameters that can be varied. Here we describe those that remained fixed in our experiments.

5.2.1. Number of samples used in CaMML's MCMC search

The default setting for these experiments is 3,515,200 samples for the MCMC search. We undertook some preliminary investigations, comparing the default with a tenth of the number of samples (351,200) and found there was no significant variation in the results. Computation time was not an issue so we ran all experiments reported below with the default value.

5.2.2. Prior arc probability

CaMML requires a probability for the existence of an arc, called P_{arc},[4] which is used in the calculations for the MML encoding. In these experiments, we set $P_{arc} = 0.5$; that is, CaMML starts by assuming there is a 50% chance there is an arc between any two variables. Expert priors can obviously override this general arc prior.

[4] If not explicitly set, CaMML estimates P_{arc} from the number of arcs found in the best model from the simulated annealing phase of its search.

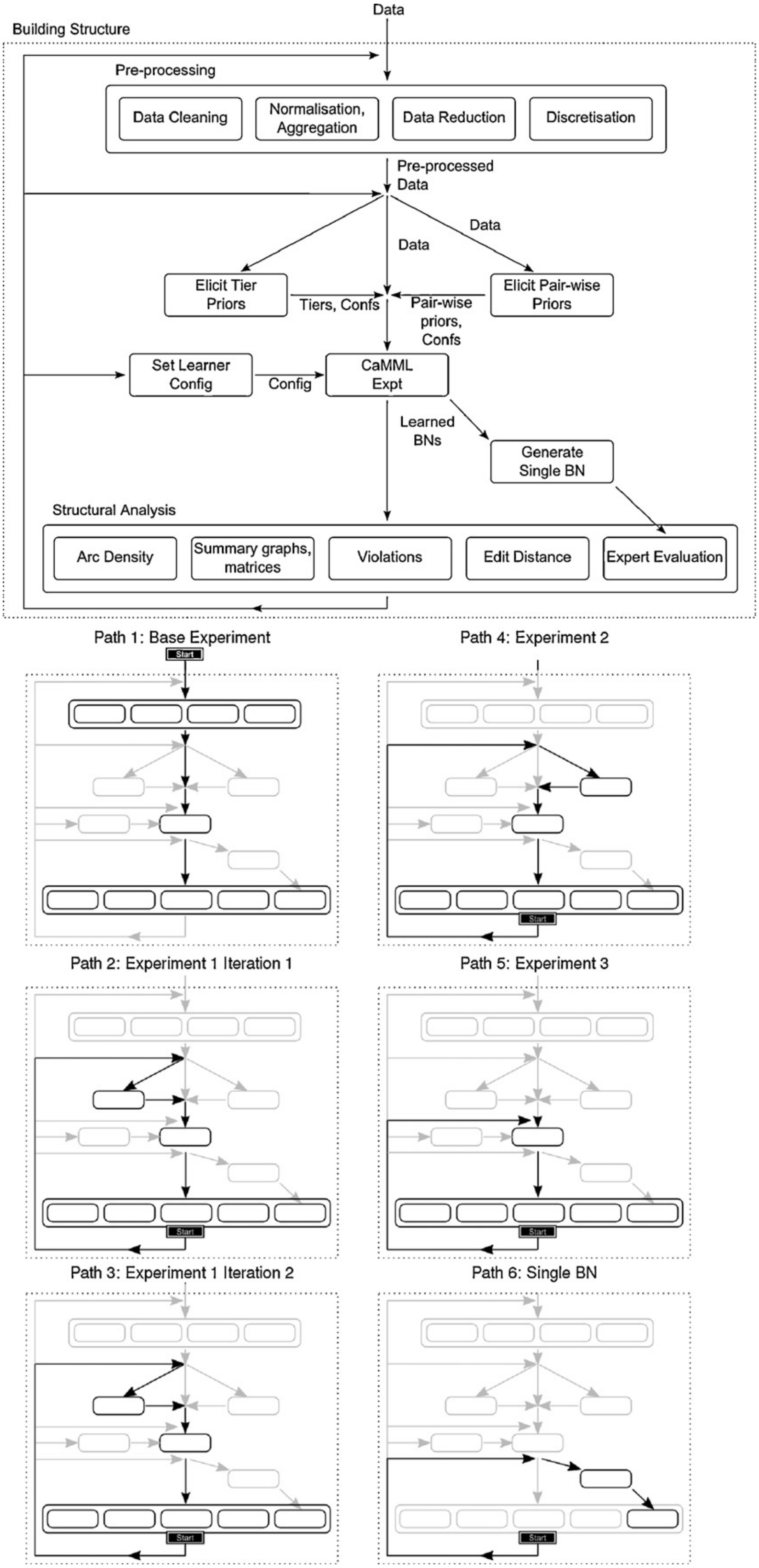


Fig. 4. Workflow showing KEBN methodology applied to dataset.

Intuitively, if the starting arc probability is low, then it takes longer and/or more data for CaMML to "find" an arc that does in fact exist.[5]

5.2.3. CaMML's output

While CaMML can produce a set of the best BNs, ranked by probability, in this paper we ran CaMML in single BN output mode.

5.2.4. Number of runs

CaMML uses a stochastic search, so we did $N = 30$ runs for each prior confidence setting to produce statistically significant results. All experimental results presented herein are based upon these 30 BNs.

5.3. Variation in structural priors

In the experiments of Section 6, we vary the structural priors in 3 ways:

1. Grouping of variables into the tiers described in Section 4.2.
   a) Original tiers: $\mathtt{bg1} \prec \mathtt{bg2} \prec \mathtt{p} \prec \mathtt{c}$. Range of confidence: 0.99, 0.9999, 0.999999, 1.
   b) Revised tiers: $(\mathtt{bg1} \cup \mathtt{bg2}) \prec \mathtt{p} \prec \mathtt{c}$. Range of confidence: 0.9, 0.99, 0.9999, 0.999999, 1.
2. Using expert (non-tier) pair-wise priors (Fig. 3), as described in Section 4.3. Range of confidence in pair-wise priors: 0.6, 0.8, 0.99.
3. Combining tier and pair-wise priors (Section 4.4), with a confidence of 0.9999 for the tier priors and varying confidence in pair-wise priors through 0.6, 0.8, 0.99.

The higher (and narrower) confidence range for the tier priors reflects the temporal basis of orderings. The lower confidence values used for the pair-wise priors reflect (i) the expert's acknowledgement that many interactions in the domain are not well understood, (ii) the expert's lack of experience in providing these pair-wise causal relationships, and (iii) the chance of individual cell errors when using the large grid format.

5.4. Evaluation measures

In BN learning research, there are a number of accepted evaluation measures. If the experiments are done on data generated from a known model, then both the structure of the learned BN and the predictions made by the learned BN, can be compared to the original model and its predictions.

However, when learning BNs using real data, as here, we cannot evaluate the resultant BN against a "gold standard". In this case, the typical approach is to learn the BN using a certain portion of the data (say 90%) and then test its predictions on the remaining 10% of the data. For each "case" (row in the data), the values of some variables are entered into the BN as evidence, belief updating is then performed, and the resultant posterior distribution for one or more target (query) variables are the model's predictions. Different measures are available to assess these predictions. A common measure is predictive accuracy: the value with the highest posterior probability is taken to be the prediction and if it matches the value in the data case, it scores 1, otherwise it scores 0. This is a very crude measure for assessing a probabilistic model, as it does not distinguish between a prediction of 0.99 (great if correct, a disaster if wrong) and 0.51 (very much sitting on the fence)—both score the same. More suitable quantitative measures include the AUC (area under the curve) of the ROC curves and Bayesian information reward [50] that also considers the calibration of the probabilistic predictions being made.

However, our focus here is to investigate the impact of structural priors on BN learning and assessing learned BN structurally as a causal model *before* parameterising the network. Predictive measures therefore are not suitable. Instead, we primarily use qualitative measures to assess and compare structures (and *only* structures) of the learned BNs.

In particular, we take CaMML's "best guess" BN from each of the 30 runs for a given configuration, and compute the following measures to evaluate and compare BN sets across different experimental configurations.

- Average arc density (# Arcs/# Nodes) and its standard deviation (s.d.);
- Average # structural prior violations;
- List actual arcs that are violations of structural priors (and the frequency with which they occurred in the 30 BNs)

Note that arc density is a commonly used metric for summarising the complexity of a BN. In general, a lower density is preferable, as a simpler network is easier to understand and requires fewer parameters. However, of course, it may be that a denser BN is a more faithful model.

5.5. Structural prior violations

Structural priors provided by the expert are violated when CaMML produces a BN that does not obey the priors (ignoring the confidence assigned to priors). An individual prior can be violated in the following ways, depending on the kind of prior:

- T he priors $\mathbf{r} \prec\prec \mathbf{c}$ and $\mathbf{r} \prec \mathbf{c}$ are violated when r is a descendant of c;
- T he priors $\mathbf{r} \rightarrow \mathbf{c}$ and $\mathbf{r} \Rightarrow \mathbf{c}$ are violated when there is no directed arc or chain from r to c;
- T he prior $\mathbf{r} - \mathbf{c}$ is violated when there is no undirected arc from r to c;
- T he prior $\mathbf{r} \sim \mathbf{c}$ is violated when there is no directed chain between r and c and no common ancestor.

5.6. Visualisation of results

5.6.1. Summary models

To aid our interpretation, we also use the arc frequency matrix produced by CaMML, depicted as a so-called weighted "summary model". This shows each arc that appeared more than a particular threshold frequency, with different line styles indicating different arc frequencies.[6] Fig. 5 shows two example summary models, showing all non-zero arc frequencies (above) and using a threshold for display of 20% frequency (below). Clearly, the 20% threshold version is significantly less cluttered. We found these summary models useful for viewing CaMML's output for a single experimental configuration, in a form that could be presented to the expert. Nevertheless, we found them cluttered, even with the 20% threshold, particularly when they included arcs in both directions between two variables (i.e., when CaMML is confident about *an* arc, but not its direction). For example, in Fig. 5(b) there are arcs in both directions between HADHRTAT and HADSTRKE.

[5] We ran some preliminary experiments with our data, varying the starting arc probability from 0.1 to 0.8 and found no significant differences in the BNs learnt. From this we concluded that we had sufficient data for this starting arc probability to be quickly overcome by the data; we therefore chose to use the value of $P_{arc} = 0.5$ in all the experiments.

[6] This format was used in [7], but there the summary was used to show frequencies in the top 10 BNs when running CaMML in the mode that produces a set of BNs for each run.

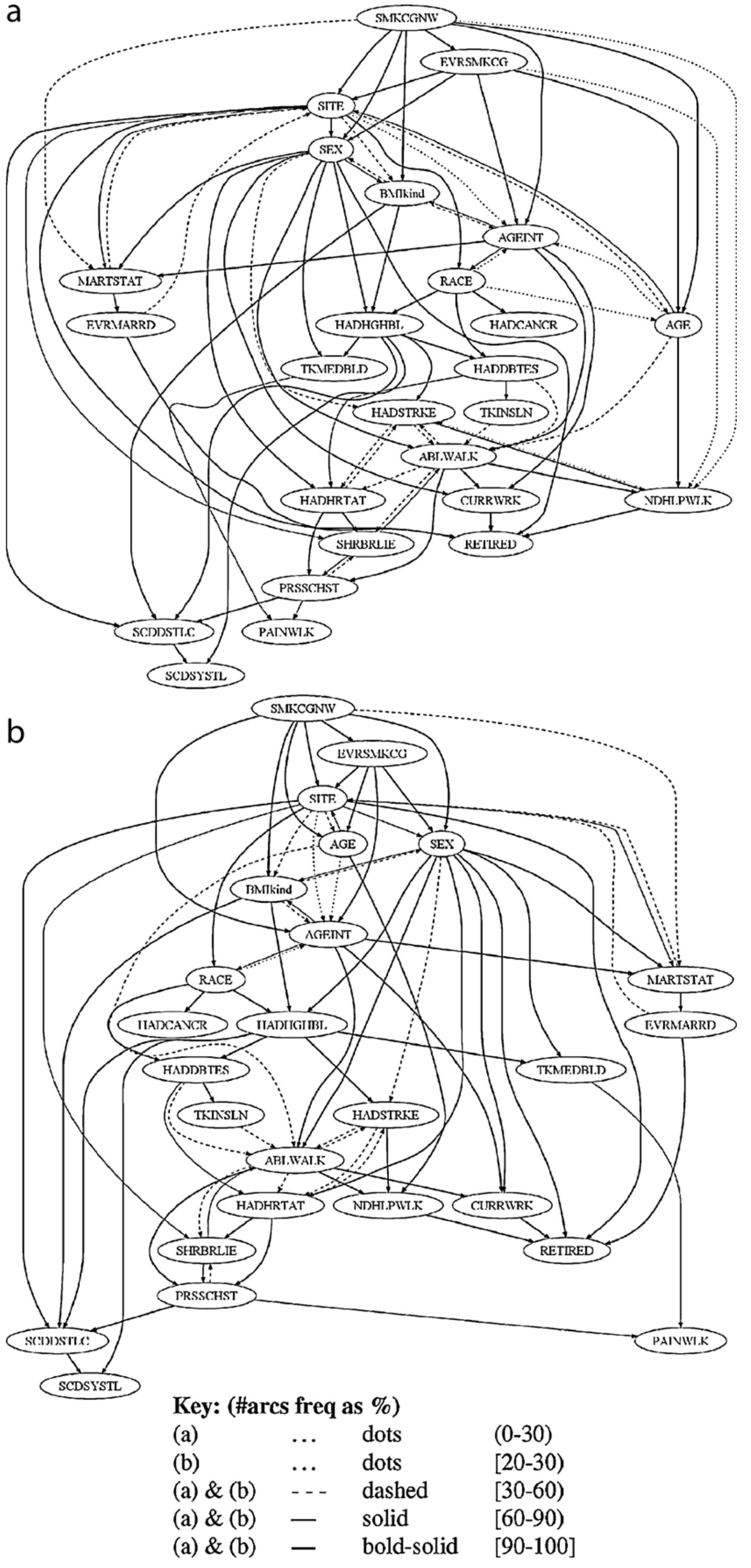


Fig. 5. Two example summary models showing arc frequencies for BNs produced by CaMML from 30 runs of one configuration (tier confidence = 0.9999, revised tiers): (a) all arc frequencies and (b) arc frequencies with the 20% threshold.

Another difficulty we encountered was that the graph layout package we used changed the layout when the arcs changed, making it very hard to compare BN structures that in fact are very similar. This made it hard to make qualitative comparisons between models produced with different configurations (e.g., different prior confidences).

5.6.2. Summary matrices

The need to evaluate the qualitative aspects of a set of learned BNs at a glance led us to develop a so-called "summary matrix", an example is given in Fig. 7. In this visualisation, the matrix shows the frequency of the arc from row to column; an entry of 1.00 indicates that arc was present in all 30 BNs, while 0.00 indicates it was present in none. We used a (green) colour gradient for the cells to reflect the arc frequencies, which draws the eye to the most frequent arcs. We note however that this format is not as useful for identifying arcs for whose direction CaMML is uncertain; For example, in Fig. 7, the opposing arcs NDHLPWLK → ABLWALK (0.30) and NDHLPWLK → ABLWALK (0.70) are both close to the diagonal and therefore each other, however others like SHRBRLIE → HADHRTAT (0.23) and HADHRTAT → SHRBRLIE (0.77) are very far apart.

We added a final feature to the summary matrix to highlight structural tier violations. Violation of a tier prior $Y \prec X$ is indicated in the cell in row Y, column X by a red rectangle overlay (for interpretation of the references to color in this figure, the reader is referred to the web version of the article). The opacity of that tier violation rectangle ranges from completely opaque (the violation occurred in 100% of the BNs) to completely transparent (the violation occurred in no BN). For example, in Fig. 7, the cell for SEX → TKMEDBLD has a red rectangle, indicating many BNs violated the tier prior SEX ≺ TKMEDBLD. In this case, the violation seems mainly due to the existence of a direct arc, TKMEDBLD → SEX, which was present in 40% of the BNs, as can be seen from the 0.40 in the cell for TKMEDBLD → SEX.

Note that it is also possible for a tier to be violated when there is no direct arc. For example the cell for NDHLPWLK → HDHRTAT has the entry 0.00, indicating no BNs contained that arc, but the partially transparent (i.e., pale pink) violation rectangle for the cell HADHRTAT → NDHLPWLK indicates that NDHLPWLK is an ancestor of HADHRTAT in at least some of the 30 BNs (though a small number). We did consider adding the actual frequency of the violation, but found no way to include it without making the summary matrix impossibly cluttered.

5.6.3. Difference matrices

Difference matrices allow us to compare the output from different experiments. They do this by showing the difference between two summary matrices, each of which summarises the 30 BNs learnt in the 30 runs for each experiment.

In these matrices, we show the variables both vertically (columns) and horizontally (rows). When the cell (i,j)[7] displays a plus symbol, the first experiment set contains more arcs of type $\mathtt{var}_i \rightarrow \mathtt{var}_j$ than the second. The minus symbol has the opposite meaning; that is, the first experiment set contains fewer arcs of type $\mathtt{var}_i \rightarrow \mathtt{var}_j$ than the second. Colouring and intensity of colour is used to indicate the intensity of this difference, with red indicating one extreme (positive/maximum) and black the other (negative/minimum). An empty (i.e., white) cell indicates that there is no difference in the number of arcs between the two sets of runs.

Fig. 8(c) shows an example difference matrix (explained in more detail in the experimental results for Experiment 1). From this matrix, we can see that the first experiment had fewer BNs with arcs from AGE → SITE, but more BNs with the arc in the opposite direction, from SITE → AGE.

5.6.4. Edit distance

Edit distance (ED) is a measure that counts the number of arcs that differ between two BNs. We use the standard three arc editing actions: addition, deletion and reversal. To produce an ED measure, our ED computation sums over the total number of edit actions needed to make one of the BNs assume the same structure as the other. For our experimental configuration, which produces a set of 30 BNs, we need to compute an average ED between the results of one configuration and another. Since performing 30 × 30 comparisons could in some sense be interpreted as double counting, our implementation takes 30 random pairings (accounting for every BN exactly once) and takes the average ED over all the pairs.

We note that the ED is a somewhat crude measure, because it fails to take into account certain qualitative information. For example, $A \rightarrow B \rightarrow C$ is 1 away from $A \rightarrow B \leftarrow C$ in ED, but represents a completely distinct set of causal relations. Nonetheless, as we will discuss in Section 6.5, this measure can give us a general understanding of learning behaviour.

5.7. Generating a single BN

While ensembles of BNs can be used for prediction or reasoning, in many applications the aim is to produce a *single* BN. Also, a number of qualitative evaluation methods (such as sensitivity analysis and scenario reviews) can be only be applied to an actual BN. Thus, the final stage of the KEBN process for our case study is to produce a single BN structure from the 30 produced by one CaMML run, for given configuration and set of priors and confidences.

We generate a single BN from the list of arc frequencies as follows (see Algorithm 5.7). Intuitively, we want to include in the single BN the arcs that appear with the highest frequency in the set. However we know that sometimes the data is such that the BN learner finds it hard to determine the direction of an arc; we have seen this uncertainty in reflected in the summary matrix by a split in the frequency, say between $A \rightarrow B$ and $B \rightarrow A$. Therefore we combine the frequency of each arc with the frequency of its reverse arc. These combined arc frequencies are then sorted from most common (i.e., appearing in 100% of generated networks) to least common. Arcs are then added to the single BN beginning with the most common, continuing down to arcs with frequencies above or equal to a user-supplied threshold. Each arc is assigned the direction that is most frequent—for example, if ABLWALK → HADSTRKE has a frequency of 40%, but the reverse arc ABLWALK → ABLWALK has a frequency of 53%, then the arc direction in the single BN is taken from the latter arc. If both directions have equal frequency, a random direction is chosen. If adding the arc produces a cycle, the reverse arc is added instead. Algorithm 1 thus produces as output a valid (i.e., acyclic) network whose arcs reflect the highest frequency arcs in the input set of BNs, down to a user specified arc frequency threshold.

The choice of threshold frequency will depend on the domain and the data, and involve at least two factors: (1) the user's preference for a simpler network with fewer arcs (high threshold) and (2) a trade-off to include as many arcs as possible that represent structural priors while not including arcs that generate prior violations. Overall, this is a more general and flexible version of Hodges et al.'s approach [51], which creates a single consensus network (but not a BN) from a set of BNs by only adding arcs with a (combined) frequency of 100%.

Another possibility is to search for the threshold that gives the "best" single structure, using any of the metrics for scoring a BN (e.g., CaMML's MML metric, the BDe [31], an MDL metric [37,38]). In Section 7 we report CaMML's MML cost metric for the single BN generated for a range of thresholds, then look in more detail at the single BNs generated using the user threshold of 70%.

[7] Here, i is the row header and j the column header.

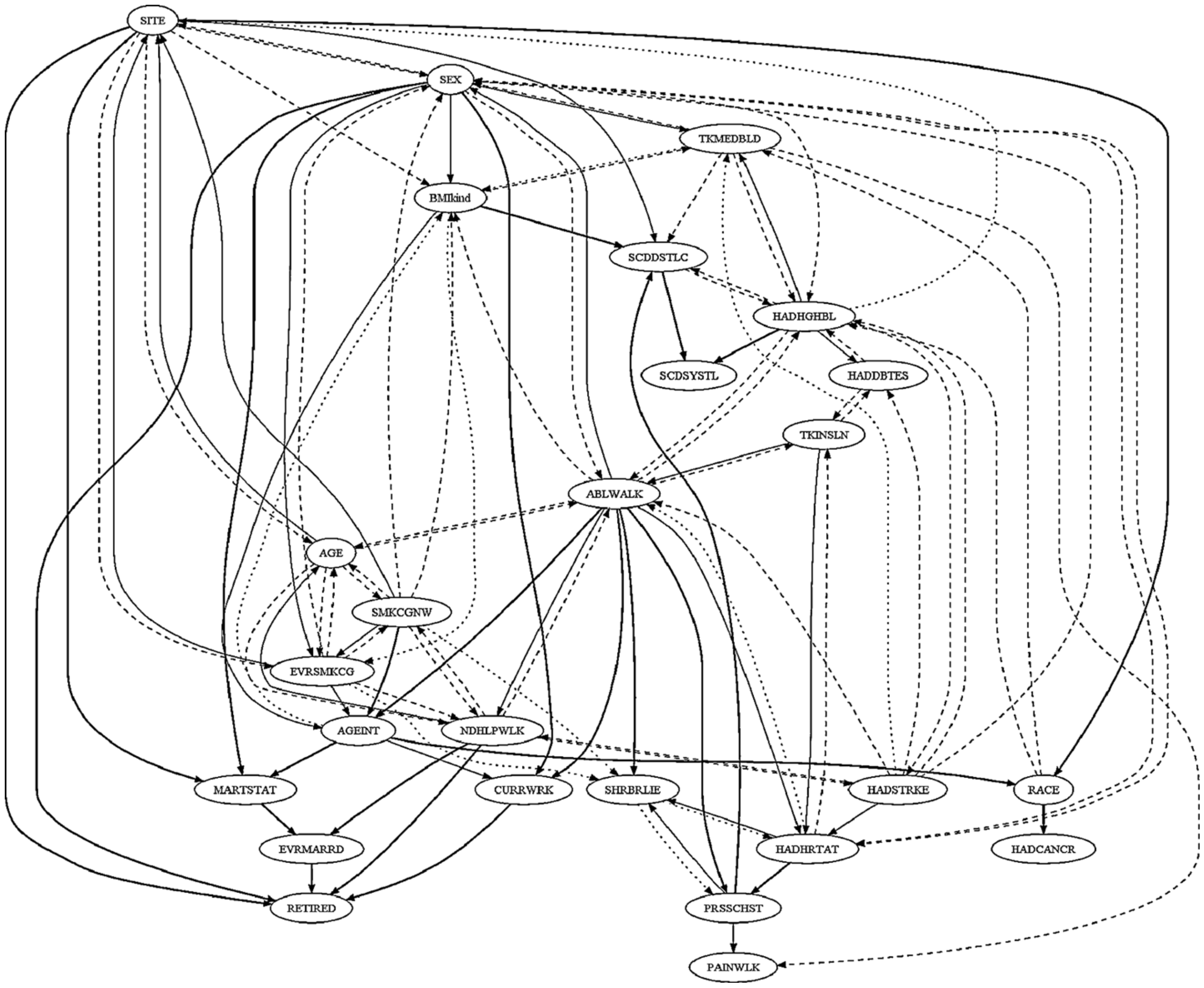


Fig. 6. Base experiment: summary model produced by CaMML from 30 runs of one configuration (arcs with freq > 0.2).

Algorithm 1 *(Generating a single BN).*

```
Input: B = a set of BNs, each containing the same set of variables V, $\tau$ = arc
      frequency threshold
Output: $G$ = a single BN, containing same variables V
Variables:
  $a$: an arc
  $f$: a frequency
  A: list of arc, frequency pairs (a, f)
A ← [] //Empty list
foreach possible arc a that exists in B do
  $f$ ← frequency ($a$ + reverse($a$) in B)
  A ← A + ($a$, $f$)
end
A ← sort A from highest frequency to lowest
$G$ ← a BN containing every variable in V and no arcs
foreach pair (a, f) in A do
  if $f < \tau$% then
    BREAK
  else if adding a to G does not produce cycle then
    add arc $a$ to $G$
  else
    add arc reverse($a$) to $G$
  end
end
```

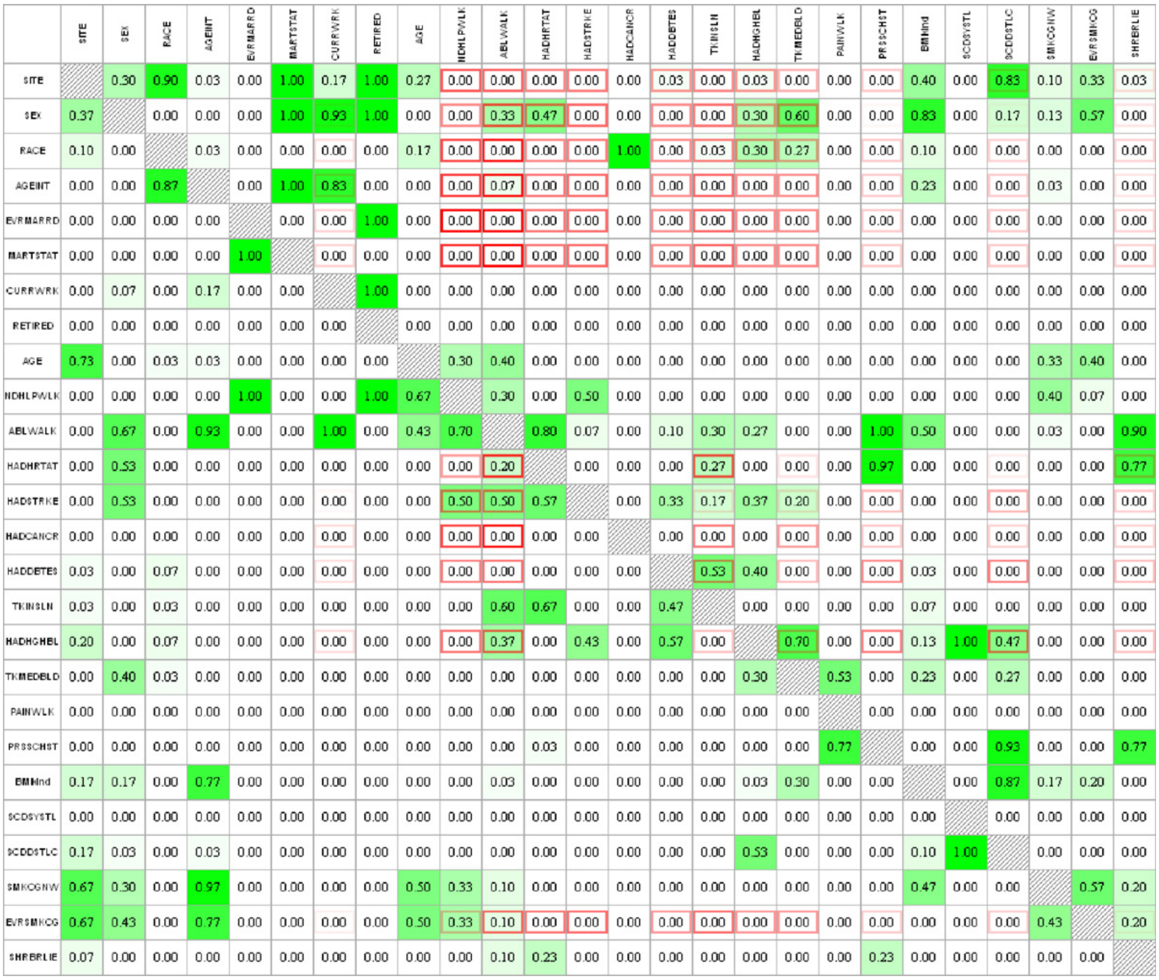
6. Experiments

6.1. Base experiment: no structural priors

Our base experiment (see workflow in Fig. 4) was to run CaMML with no structural priors, providing a base against which to compare the results from our subsequent experiments with structural priors.

The average arc density of the learned BNs is 2.46 (with s.d. 0.0655), indicating that CaMML learns a relatively sparse network from the data only. This corresponded to the average number of arcs being about 64. This suggests that the 84 causal relationships → (which correspond to an arc in the BN), specified by our expert, may not be supported by the data.

The summary model is given in Fig. 6, with the corresponding summary matrix shown in Fig. 7. Clearly, there are many low frequency arcs and very few "certain arcs"; for example, only 15 arcs with frequency 1.00. CaMML is also often uncertain of the arc direction; there are 24 "pairs" of nodes X,Y where both $X \rightarrow Y$ and $Y \rightarrow X$ have frequency > 0.2 (the threshold used for including in the summary model in Fig. 6). This shows how difficult it is for a BN learner,



	SITE	SEX	RACE	AGEINT	EVRMARRD	MARTSTAT	CURRWRK	RETIRED	AGE	NDHLPWLK	ABLWALK	HADHRTAT	HADSTRKE	HADCANCR	HADDBTES	TKINSLN	HADHGHBL	TKMEDBLD	PAINWLK	PRSSCHST	BMIind	SCDSYSTL	SCDDSTLC	SMKOGNW	EVRSMKOG	SHRBRLIE
SITE		0.30	0.90	0.03	0.00	1.00	0.17	1.00	0.27	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.00	0.00	0.40	0.00	0.83	0.10	0.33	0.03
SEX	0.37		0.00	0.00	0.00	1.00	0.93	1.00	0.00	0.00	0.33	0.47	0.00	0.00	0.00	0.00	0.30	0.60	0.00	0.00	0.83	0.00	0.17	0.13	0.57	0.00
RACE	0.10	0.00		0.03	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	1.00	0.00	0.03	0.30	0.27	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
AGEINT	0.00	0.00	0.87		0.00	1.00	0.83	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.03	0.00	0.00
EVRMARRD	0.00	0.00	0.00	0.00		0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MARTSTAT	0.00	0.00	0.00	0.00	1.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CURRWRK	0.00	0.07	0.00	0.17	0.00	0.00		1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RETIRED	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGE	0.73	0.00	0.03	0.03	0.00	0.00	0.00	0.00		0.30	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.40	0.00
NDHLPWLK	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.67		0.30	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.07	0.00
ABLWALK	0.00	0.67	0.00	0.93	0.00	0.00	1.00	0.00	0.43	0.70		0.80	0.07	0.00	0.10	0.30	0.27	0.00	0.00	1.00	0.50	0.00	0.00	0.03	0.00	0.90
HADHRTAT	0.00	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20		0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.77
HADSTRKE	0.00	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.57		0.00	0.33	0.17	0.37	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HADCANCR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HADDBTES	0.03	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.53	0.40	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
TKINSLN	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.67	0.00	0.00	0.47		0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00
HADHGHBL	0.20	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.00	0.43	0.00	0.57	0.00		0.70	0.00	0.00	0.13	1.00	0.47	0.00	0.00	0.00
TKMEDBLD	0.00	0.40	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30		0.53	0.00	0.23	0.00	0.27	0.00	0.00	0.00
PAINWLK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00
PRSSCHST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.77		0.00	0.00	0.93	0.00	0.00	0.77
BMIind	0.17	0.17	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.30	0.00	0.00		0.00	0.87	0.17	0.20	0.00
SCDSYSTL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00
SCDDSTLC	0.17	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.53	0.00	0.00	0.00	0.10	1.00		0.00	0.00	0.00
SMKOGNW	0.67	0.30	0.00	0.97	0.00	0.00	0.00	0.00	0.50	0.33	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.00	0.00		0.57	0.20
EVRSMKOG	0.67	0.43	0.00	0.77	0.00	0.00	0.00	0.00	0.50	0.33	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43		0.20
SHRBRLIE	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.00	

Fig. 7. Base experiment: summary matrix produced by CaMML with no structural priors.

even with what would be considered a reasonable amount of data, to learn a BN with this many variables.

For the purpose of later comparisons, we note that the average number of tier prior violations in these BNs (for the original tiers) is 48.03 (with s.d. 16.93), while the average number of violations of the expert pair-wise priors is 71.23 (with s.d. 2.87) (although of course the priors were not provided to the CaMML learning algorithm in this experiment). This suggests, even before we run experiments incorporating them, that the tier priors agree more with the data than the pair-wise priors.

Overall, the base experiment confirms the premise underpinning this research: that often BN learners find it difficult to learn useful models when there are large numbers of variables involved. In this case, the resulting networks are complex, with a lot of uncertainty about the structure, and even ignoring the weakest arcs (as in Fig. 5(a) and (b)), make no sense as causal models of the domain.

6.2. Experiment 1: expert tier structural priors

In this experiment we incorporated tier priors, following the QMR model (Section 2.2.1) and our analysis of temporal aspects of the problem (described in Section 4.2). The experiment was performed in two iterations:

[Iteration 1:] Original tiers, over confidence values {0.99, 0.9999, 0.999999, 1}.

[Iteration 2:] Revised tiers, over confidence values {0.9, 0.99, 0.9999, 0.999999, 1}.

with the tiers revised for Iteration 2 after analysis of results from Iteration 1 (see below).

The arc density (see Fig. 8(a)), which represents the mean number of arcs per node, varies between roughly 2.3 and 2.4 with only the value for tier prior confidence value 1 (original tiers) being statistically significantly different (lower) than the others. The arc densities from the revised tier runs with confidence 0.9, 0.9999999 and 1.0 are also significantly lower than the arc density for the base (no priors) experiment. These results indicate that incorporating priors *may* result in sparser (and simpler) models, even though the original model was fairly compact anyway.

Fig. 8(b) shows the average number of structural tier violations. In all cases the number of violations is always under 10, and far fewer than the 48.03 in the base experiment. Adding the tier priors makes a substantial difference in the structures CaMML learns. The higher the confidence level in the tier priors, the fewer tier violations in the learned networks, (although there is no significant difference in violations for the 0.9999 and 0.999999 confidence). This is exactly what we would expect; by giving higher confidence in the tier priors, CaMML is giving greater weight to the prior, so it is harder for the data to overcome the tiers, and hence there are fewer violations. Also, for each confidence level, the revised tiers produced fewer violations. Again, this is expected, as tier priors

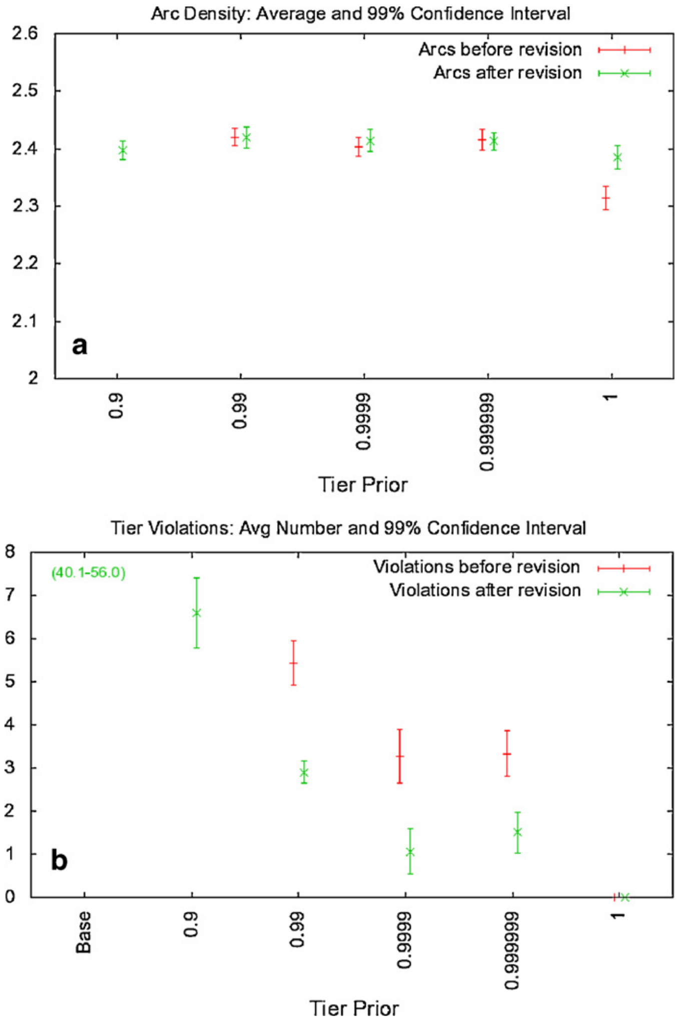


Fig. 8. Experiment 1: original tiers vs revised tiers, varying confidence 0.9, 0.99, 0.9999, 0.999999, 1, (a) average arc density (b) average no. of structural tier violations (both with 99% confidence interval).

that lead to violations in the original tiers used in Iteration 1 were removed for the Iteration 2 (revised tiers).

Table 2 shows a list of the tier priors violated, for the two sets of priors. There were two main types of violations. First, there were many that were of the form $\mathtt{bg}_2 \rightarrow \mathtt{bg}_1$, violating $\mathtt{bg}_1 \prec \mathtt{bg}_2$. This indicated that the distinction we had made between different kinds of background factors (that is, splitting bg into $\mathtt{bg}_1$ and $\mathtt{bg}_2$) was not warranted, and hence motivated the removal of this distinction in our revised tiers.

The second type of violation involved interactions between “past” variables (e.g., HADSTRKE and HADHRTAT), and variables that we considered indicators of current health (NDHLPWLK, ABLWALK, TKINSLN). We postulate that these tier violations are due to CaMML having difficulties in getting arc directions correct when the important variables are hidden. For example, if the true actual structure is that of Fig. 10, then if the *Disease* nodes are hidden (i.e., not in the dataset), tier priors $symptom_t \prec symptom_{t+1}$ with anything less than confidence 1 may not be enough for CaMML to learn $symptom_t \rightarrow symptom_{t+1}$. We note that with the revised tiers, these violations no longer occurred (even though there were no changes to the tiers involving these variable), and those remaining occurred much less often.

We also note that, for the lower confidence of 0.9, the revised tiers resulted in a set of tier violations involving EVRMARRD and two variables from the ‘past’ (p) tier and again two from the ‘current indirect’ (ci) tier. They all occurred just over half the time (frequency 0.57).

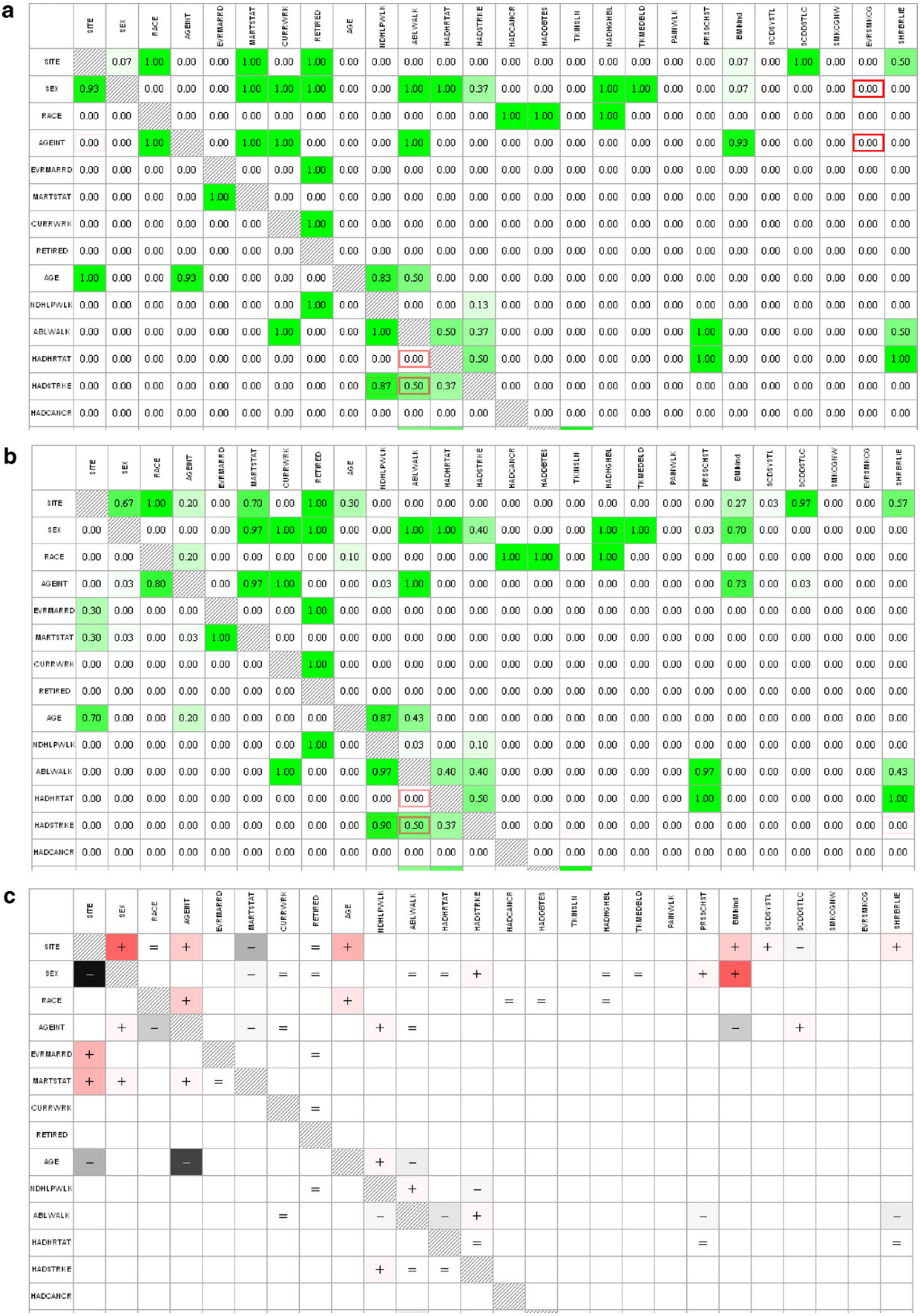
In the summary matrices for both iterations (Fig. 9(a) and (b), top halves only shown due to space), now, in contrast to the base experiment, there are more “extreme” values in the edge frequencies with many 1.0 (present in all BNs) and 0.0 (absent in all BNs) and fewer “weak” (i.e., less frequent) edges. We saw above in Fig. 8 that the *average* number of tier prior violations for confidence 0.9999 dropped from 3 for the original priors to 1 for the revised priors. The summary matrices show visually *which* tier priors are violated and *how often*: 7 with the original priors, 5 for the revised priors (listed in Table 2). Note that those with frequency less than 0.1 are too faint to see clearly at this resolution.

Fig. 9(c) shows the difference matrix for the original and revised priors. Most changes are minor (depicted by the pale pink and pale grey). The use of intensity allows us to see quickly that the main changes are in the links between SITE-SEX, SEX-BMIKind and AGE-AGENT. These are all background variables, which were in the tier priors we changed, so the reduction in the violations has been achieved with minimal change to the network structure as a whole.

Finally, Fig. 12 compares all four experiments – Base, 1, 2 and 3 – using differences matrices between each pair of experiments. Recall from Section 5.6 that a difference matrix summarises which arcs are the same (white), present more frequently (+ sign, redder the higher) and present less frequently (– sign, darker the fewer). While obviously the size of the matrices in this figure does not allow a detailed analysis, it does allow us to identify general trends in the

Table 2
Experiment 1: listing of tier violations.

Tier	Expt 1 (original tiers)			Expt 1 (revised tiers)			
	0.99	0.9999	0.999999	0.9	0.99	0.9999	0.999999
SEX ≺ SITE	0.40	0.07					
SEX ≺ EVRSMKCG	1.00	1.00	1.0				
AGEINT ≺ SITE	0.30	0.07					
AGEINT ≺ EVRSMKCG	1.00	1.00	1.00				
HADHRTAT ≺ ABLWALK	0.97	0.50	0.63	1.00	1.00	0.40	0.70
HADHRTAT ≺ TKINSL	0.60			1.00	0.63		
HADSTRKE ≺ NDHLPWLK	0.20	0.13	0.13	0.97	0.23	0.10	0.10
HADSTRKE ≺ ABLWALK	0.97	0.50	0.57	1.00	1.00	0.47	0.70
HADHRTAT ≺ SHRBRLIE				0.03			
HADSTRKE ≺ SHRBRLIE				0.03		0.07	
HADSTRKE ≺ TKINSLN				0.30	0.03	0.03	
EVRMARRD ≺ HADHGHB				0.57			
EVRMARRD ≺ ABLWALK				0.57			
EVRMARRD ≺ NDHLPWLK				0.57			
EVRMARRD ≺ HADDBTES				0.57			



a

	SITE	SEX	RACE	AGEINT	EVRMARRD	MARTSTAT	CURRWRK	RETIRED	AGE	NDHLPWLK	ABLWALK	HADHRTAT	HADSTRKE	HADCANCR	HADDBTES	TKINSLN	HADHGHBL	TKMEDBLD	PAINWLK	PRSSCHST	EMHind	SCDSYSTL	SCDDSTLC	SMKCGNW	EVRSMKCG	SHRBRLIE
SITE	▨	0.07	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	1.00	0.00	0.00	0.50
SEX	0.93	▨	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.37	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00
RACE	0.00	0.00	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGEINT	0.00	0.00	1.00	▨	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00
EVRMARRD	0.00	0.00	0.00	0.00	▨	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MARTSTAT	0.00	0.00	0.00	0.00	1.00	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CURRWRK	0.00	0.00	0.00	0.00	0.00	0.00	▨	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RETIRED	0.00	0.00	0.00	0.00	0.00	0.00	0.00	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGE	1.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	▨	0.83	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NDHLPWLK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	▨	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABLWALK	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	▨	0.50	0.37	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.50
HADHRTAT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	▨	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
HADSTRKE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.50	0.37	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HADCANCR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

b

	SITE	SEX	RACE	AGEINT	EVRMARRD	MARTSTAT	CURRWRK	RETIRED	AGE	NDHLPWLK	ABLWALK	HADHRTAT	HADSTRKE	HADCANCR	HADDBTES	TKINSLN	HADHGHBL	TKMEDBLD	PAINWLK	PRSSCHST	EMHind	SCDSYSTL	SCDDSTLC	SMKCGNW	EVRSMKCG	SHRBRLIE
SITE	▨	0.67	1.00	0.20	0.00	0.70	0.00	1.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.03	0.97	0.00	0.00	0.57
SEX	0.00	▨	0.00	0.00	0.00	0.97	1.00	1.00	0.00	0.00	1.00	1.00	0.40	0.00	0.00	0.00	1.00	1.00	0.00	0.03	0.70	0.00	0.00	0.00	0.00	0.00
RACE	0.00	0.00	▨	0.20	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGEINT	0.00	0.03	0.80	▨	0.00	0.97	1.00	0.00	0.00	0.03	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.00	0.03	0.00	0.00	0.00
EVRMARRD	0.30	0.00	0.00	0.00	▨	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MARTSTAT	0.30	0.03	0.00	0.03	1.00	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CURRWRK	0.00	0.00	0.00	0.00	0.00	0.00	▨	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RETIRED	0.00	0.00	0.00	0.00	0.00	0.00	0.00	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGE	0.70	0.00	0.00	0.20	0.00	0.00	0.00	0.00	▨	0.87	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NDHLPWLK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	▨	0.03	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABLWALK	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.97	▨	0.40	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.43
HADHRTAT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	▨	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
HADSTRKE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.50	0.37	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HADCANCR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	▨	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

c

	SITE	SEX	RACE	AGEINT	EVRMARRD	MARTSTAT	CURRWRK	RETIRED	AGE	NDHLPWLK	ABLWALK	HADHRTAT	HADSTRKE	HADCANCR	HADDBTES	TKINSLN	HADHGHBL	TKMEDBLD	PAINWLK	PRSSCHST	EMHind	SCDSYSTL	SCDDSTLC	SMKCGNW	EVRSMKCG	SHRBRLIE
SITE	▨	+	=	+		−		=	+												+	+	−			+
SEX	−	▨				−	=	=			=	=	+				=	=		+	+					
RACE			▨	+					+					=	=		=									
AGEINT		+	−	▨		−	=			+	=										−		+			
EVRMARRD	+				▨			=																		
MARTSTAT	+	+		+	=	▨																				
CURRWRK							▨	=																		
RETIRED								▨																		
AGE	−			−					▨	+	−															
NDHLPWLK								=		▨	+		−													
ABLWALK							=			−	▨	−	+							−						−
HADHRTAT												▨	=							=						=
HADSTRKE										+	=	=	▨													
HADCANCR														▨												

Fig. 9. Experiment 1 (a) Iteration 1: summary matrix (top half) from original priors, (b) Iteration 2: summary matrix (top half) from revised priors, (c) difference matrix (top half) comparing the revised priors with the original priors. All confidence 0.9999.

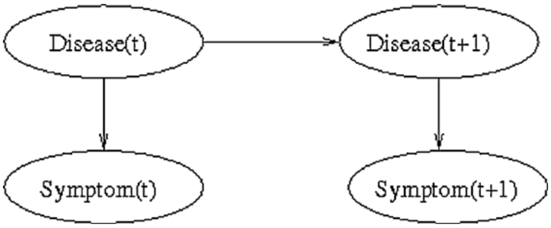


Fig. 10. BN fragment showing possible underlying “true” structure; $symptom_t \rightarrow symptom_{t+1}$ should be learnt when Disease variables are unobserved.

comparative results. For example, the tier priors in Experiment 1 (0.9999) clearly give a summary matrix that is quite different to that learnt in the Base Experiment, with a number of darker red and black cells and many pale coloured cells. The symmetry of large parts of the difference matrix suggests that many of the changes relate to change in arc direction rather than arc existence. (We'll consider other aspects of this figure as we look at the results for each experiment in turn.)

6.3. Experiment 2: expert pair-wise structural priors

In this experiment we provide CaMML with the expert's pair-wise (non-tier) priors (Fig. 3), with confidence values {0.6, 0.8, 0.9}. First, we can see from Fig. 11(a) that the average arc density of the Experiment 2 networks is similar to those of the Base and Experiment 1, ranging from 2.40 (99% confidence interval [2.37, 2.43]) for confidence 0.6–2.45 (confidence interval [2.42, 2.48]) for confidence 0.8. The slightly higher density for the higher confidences, though not statistically significant, indicates that when the pair-wise priors are given more weight, the results BNs are slightly more complex.

In this experiment, when looking at the average number of prior violations, we distinguish between violations of the pair-wise priors (Fig. 11(b)) and violations of structural tier priors (Fig. 11(c)). The number of violations (of both pair-wise and tier priors) for the lowest pair-wise prior confidence (0.6) is indistinguishable from the Base Experiment, suggesting that such a low confidence is not enough to influence CaMML's search, given the amount of data available. We can see, however, that providing CaMML with the stronger pair-wise priors decreases the average number of pair-wise prior violations. It is striking that, even with the highest confidence (0.99) there are still many violations (about 67). This suggests that the set of pair-wise priors, as a whole, are not in agreement with the data. Increasing the confidence in the pair-wise priors also resulted in a slight reduction in the number of tier prior violations, suggesting some overlap in the relationships being represented by the two types of structural priors.

Fig. 13(a) shows the summary matrix for the 0.8 confidence run (the matrices for 0.6 and 0.99 are not included for reasons of space). Looking first at the arcs that CaMML is “certain” about (i.e., where the arc frequency is 1.00), we see there are only 12 such arcs, less than 19% of the average 64 arcs. There are also few high frequency arcs; only 6 arcs have frequency between 0.90 and 1.00.

Fig. 13(a) also shows which pair-wise priors are violated in the learned BNs, with the intensity of the red ovals indicating the frequency. While the violation visualisation does not distinguish between types of relationships (e.g., –– vs →), it was straightforward to overlay the violations on the original priors (though not included here). This showed that the vast majority of the violations are the direct causal pairwise relationships (→).

From Fig. 12 we can see that the pair-wise priors (0.8) make relatively few changes to the Base Experiment matrix, compared to the impact of tier priors in Experiment 1. The difference matrices between Experiments 1 and 2, on the other hand, show far more differences. In many cases, strong differences in cells between Base to Experiment 1 (say red) are reversed in the difference matrix cells between Experiments 1 and 2.

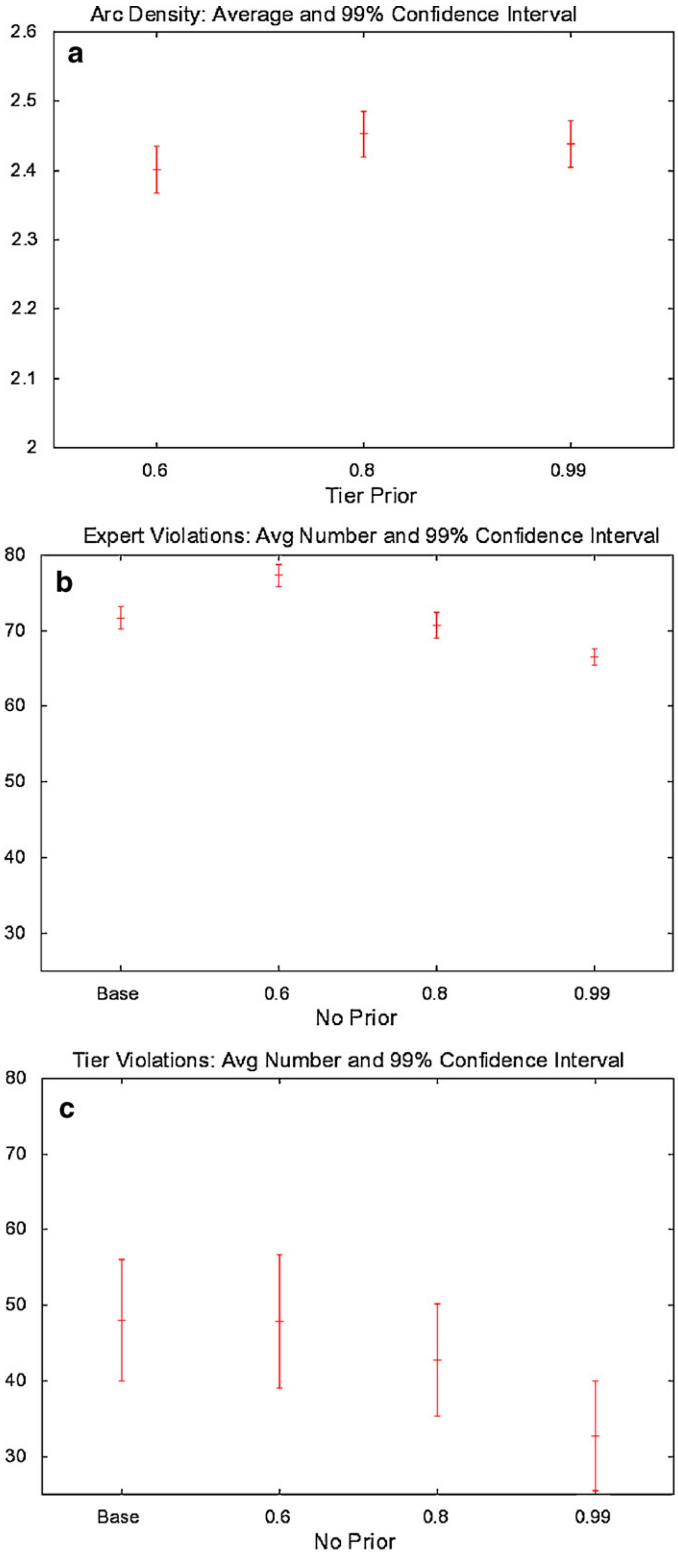


Fig. 11. Experiment 2: learning with expert pair-wise priors, varying confidence 0.6, 0.8, 0.99. (a) Average arc density, (b) average no. of violations of pair-wise priors, (c) average no. of tier prior violations. (All with 99% confidence interval.)

In summary, for this experiment with pair-wise priors only, there are large number of weak arcs in the summary matrix, a limited reduction in the number of pair-wise prior violations, relatively

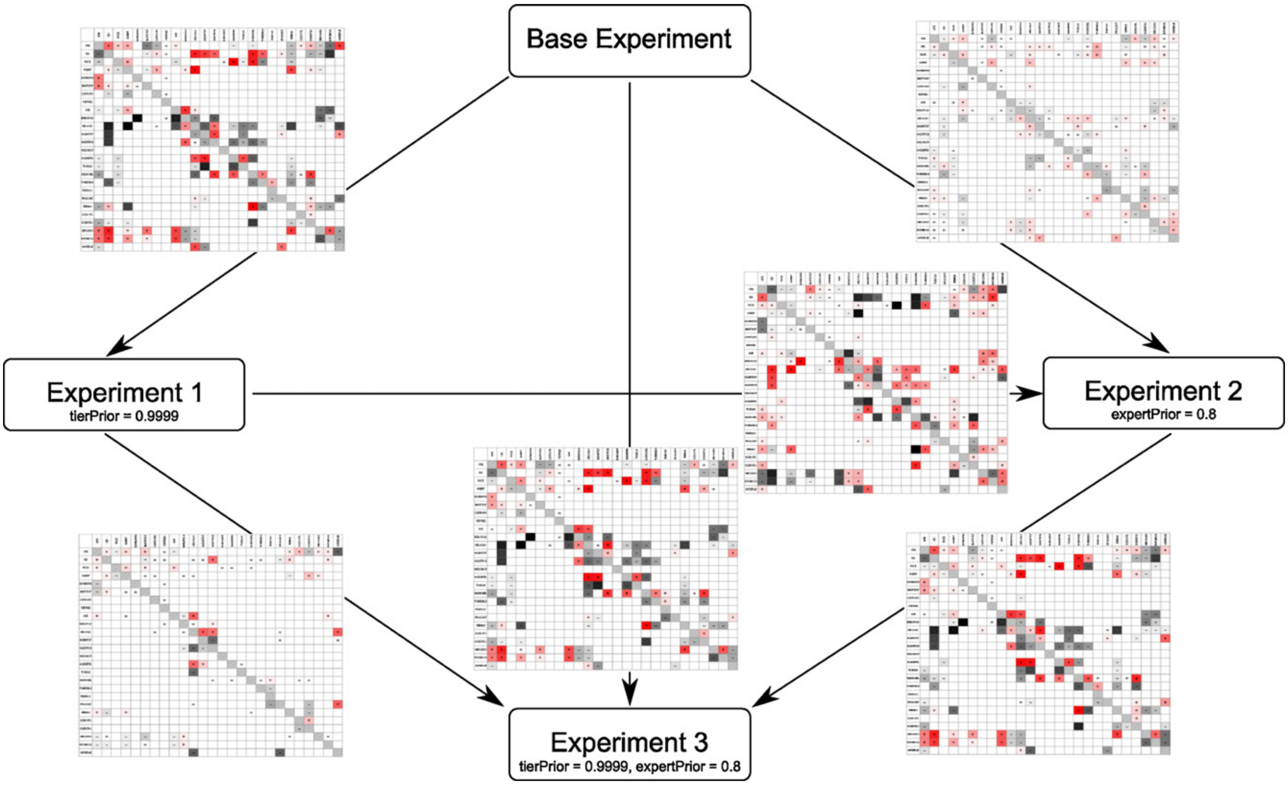


Fig. 12. Difference matrices across whole experiment workflow.

few changes from the base case and still many violations, especially of the direct → priors. All this suggests that the pair-wise priors do not fit the data. Clearly, despite the iterative communications with the expert, too many and too *fine* pairwise relationships were elicited.

Rather than continue iterating with variations in the pair-wise priors (which, of course, is also a reasonable option), for the purposes of this study we chose to keep the original, obviously flawed set, and see what happened when we combined them with the tier priors.

6.4. Experiment 3: combining tier, pair-wise priors

In this experiment, we combine the tier and arc priors and compare the results with those of the two previous experiments, where they were applied separately. We use a single confidence 0.9999 for the revised tier priors, but vary the confidence in the pair-wise priors over the same range, namely 0.6, 0.8 and 0.99.

As we would expect, as the confidence in the pair-wise priors goes up, the average number of pair-wise prior violations (Fig. 14(b)) goes down, while on average the number of tier prior violations (Fig. 14(b)) goes up (although the only statistically significant differences are between the lowest value and the others).

The average arc density results for Experiment 3 are shown in Fig. 14(a). While a confidence of 0.8 and 0.99 for the pair-wise priors resulted in a similar average arc density to the previous experiments (around 2.4), interestingly, a confidence of 0.6 gave a significantly lower average arc density, yielding the confidence interval [2.34, 2.38]. This is also significantly lower than the average arc density obtained with tier priors with confidence 0.9999 (Experiment 1, Fig. 8(a)).

Looking at the summary matrix from Experiment 3 (confidence 0.8) (Fig. 13(b)) we can see that there are more "certain" arcs (cells with 1.00) than compared to Experiment 2, but slightly fewer than for Experiment 1. We note that all the tier prior violations involve `ABLWALK` (in the second position in the pair-wise relationships, i.e., X ≺ `ABLWALK`). Certainly not all strokes lead causally to persistent failure to walk,[8] however there is no obvious causal sequence in the opposite direction. We note however that the data contains no time marks and hence gives no indication as to how longstanding any walking ability difficult may be, and whether it pre-dates any past stroke; violation of this tier prior is therefore unsurprising. Also, all of the remaining pair-wise prior violations involve the → relationship given by the expert; the reduction in pair-wise prior violations has been large in terms of the other relationships.

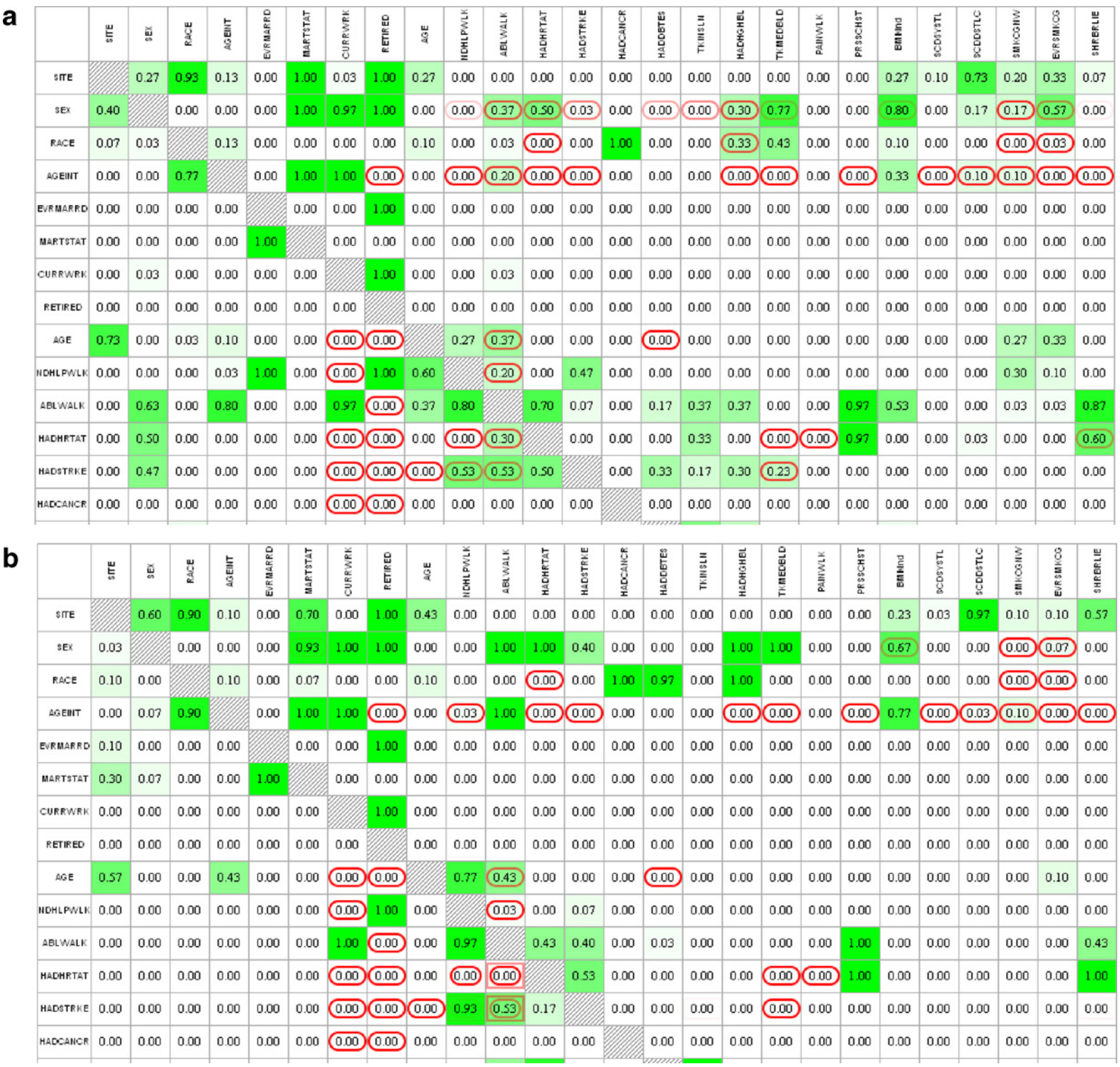
Finally, the difference matrices between Experiment 3 and the other preceding experiments (Fig. 12) show the 3 different "paths" to the same outcome. The combined priors have resulted in only slight differences, compared to tiers only. It's clear that tier priors in this case are providing the major assistance to causal discovery and that arc priors have only had a modest impact.

6.5. Edit distance comparison across experiments

As described in Section 5.4, we can look at the ED between the 30 BNs produced for each configuration of each experiment. Fig. 15 shows for each pair of experiments, the average ED and the standard deviation, in figures, with the shading giving a visualisation of the ED (darker means higher average ED). The darker lines across the grid delineate the different experiments, with each row and column showing the confidence in the prior used.

The edit distances between the Base Experiment and all other experiments (first column) are all large (over 30), which we might expect—the priors are making a difference to the networks learner. The most difference is between the Base and the combined priors

[8] And, of course difficulty walking may be due to non stroke causes such as arthritis or indeed congestive heart failure.



a

	SITE	SEX	RACE	AGEINT	EVRMARRD	MARTSTAT	CURRWRK	RETIRED	AGE	NDHLPWLK	ABLWALK	HADHRTAT	HADSTRKE	HADCANCR	HADDBTES	TKINSLN	HADHGHBL	TKMEDBLD	PAINWLK	PRSSCHST	BMInd	SCDSYSTL	SCDDSTLC	SMKCGNW	EVRSMKCG	SHRBRLIE
SITE		0.27	0.93	0.13	0.00	1.00	0.03	1.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.10	0.73	0.20	0.33	0.07
SEX	0.40		0.00	0.00	0.00	1.00	0.97	1.00	0.00	0.00	0.37	0.50	0.03	0.00	0.00	0.00	0.30	0.77	0.00	0.00	0.80	0.00	0.17	0.17	0.57	0.00
RACE	0.07	0.03		0.13	0.00	0.00	0.00	0.00	0.10	0.00	0.03	0.00	0.00	1.00	0.00	0.00	0.33	0.43	0.00	0.00	0.10	0.00	0.00	0.00	0.03	0.00
AGEINT	0.00	0.00	0.77		0.00	1.00	1.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.10	0.10	0.00	0.00
EVRMARRD	0.00	0.00	0.00	0.00		0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MARTSTAT	0.00	0.00	0.00	0.00	1.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CURRWRK	0.00	0.03	0.00	0.00	0.00	0.00		1.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RETIRED	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGE	0.73	0.00	0.03	0.10	0.00	0.00	0.00	0.00		0.27	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.33	0.00
NDHLPWLK	0.00	0.00	0.00	0.03	1.00	0.00	0.00	1.00	0.60		0.20	0.00	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.10	0.00
ABLWALK	0.00	0.63	0.00	0.80	0.00	0.00	0.97	0.00	0.37	0.80		0.70	0.07	0.00	0.17	0.37	0.37	0.00	0.00	0.97	0.53	0.00	0.00	0.03	0.03	0.87
HADHRTAT	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30		0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.97	0.00	0.00	0.03	0.00	0.00	0.60
HADSTRKE	0.00	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.53	0.53	0.50		0.00	0.33	0.17	0.30	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HADCANCR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

b

	SITE	SEX	RACE	AGEINT	EVRMARRD	MARTSTAT	CURRWRK	RETIRED	AGE	NDHLPWLK	ABLWALK	HADHRTAT	HADSTRKE	HADCANCR	HADDBTES	TKINSLN	HADHGHBL	TKMEDBLD	PAINWLK	PRSSCHST	BMInd	SCDSYSTL	SCDDSTLC	SMKCGNW	EVRSMKCG	SHRBRLIE
SITE		0.60	0.90	0.10	0.00	0.70	0.00	1.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.03	0.97	0.10	0.10	0.57
SEX	0.03		0.00	0.00	0.00	0.93	1.00	1.00	0.00	0.00	1.00	1.00	0.40	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.67	0.00	0.00	0.00	0.07	0.00
RACE	0.10	0.00		0.10	0.00	0.07	0.00	0.00	0.10	0.00	0.00	0.00	0.00	1.00	0.97	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGEINT	0.00	0.07	0.90		0.00	1.00	1.00	0.00	0.00	0.03	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.00	0.03	0.10	0.00	0.00
EVRMARRD	0.10	0.00	0.00	0.00		0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MARTSTAT	0.30	0.07	0.00	0.00	1.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CURRWRK	0.00	0.00	0.00	0.00	0.00	0.00		1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RETIRED	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGE	0.57	0.00	0.00	0.43	0.00	0.00	0.00	0.00		0.77	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
NDHLPWLK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00		0.03	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABLWALK	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.97		0.43	0.40	0.00	0.03	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.43
HADHRTAT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.53	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
HADSTRKE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.53	0.17		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HADCANCR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Fig. 13. Summary matrices (top-half) (a) Experiment 2, pair-wise prior confidence = 0.8 (b) Experiment 3, revised tiers, tier confidence = 0.9999, expert prior confidence = 0.8. Violations of pair-wise priors indicated by red ovals, tier prior violation indicated by red rectangle, intensity indicates frequency of violations.

(with no statistically significant difference when varying the pair-wise prior confidence). Otherwise, the bigger edit distances are found between the Experiment 2 (pair-wise prior only experiments) and both experiments 1 and 3, reconfirming the distinction between these indicated by the difference matrices.

Within an experiment, increasing the confidence sometimes increased the average ED; for example, for when comparing Experiment 1 with itself (column conf = 0.9), as the tier prior increased from 0.99 to 1, the average ED increases from 14.6 to 19.5. On the other hand, in some cases (e.g., Experiment 3, conf = 0.6, when compared to Experiment 1), increasing the Experiment 1 confidence from 0.9 to 0.9999 (or higher) decreases the average ED; this is because the run becomes closer to Experiment 3 with tier prior conf = 0.9999.

7. Generating a single BN

As described in Section 5.7, the final stage of the KEBN process for our case study was to generate a single BN structure, using a threshold frequency, from the 30 produced by one CaMML run. We performed this procedure with thresholds from 100% down to 0%, using 1% steps, for Experiment 1 (tier conf = 0.9999) and Experiment 3 (tier conf = 0.9999, pair-wise conf = 0.8). Fig. 16 shows both the number of arcs in the single BN generated for each threshold, and the MML cost for each structure. We can see that the steps are coarser than the 1% threshold step, as there are no arcs at many frequencies. The MML score decreases for a while as the threshold descrease,[9] then increases as less frequent arcs are added.

Fig. 17 shows the single BN structure generated in Experiment 3 for the threshold frequency 70%, an example of a threshold that gives a BN slightly simpler than the "best" BN according to the MML cost in each case. As mentioned in Section 5.7, the appropriate trade-off will be user- and domain-specific. The unlabelled arcs are those in the single BN generated with 100% threshold frequency, arcs labelled B show the first set of arcs added (with frequency 96.67%), and so on for C (86.66%) including ABLWALK → HADHRTAT which violates tier prior $p \prec c$, D (83.33%), E

[9] It is interesting to see that the MML cost varies by only about 1–2% across all frequencies. This is because the large amount of data for this case study means that the cost for representing the data given the BN far outweighs the cost for the BN.

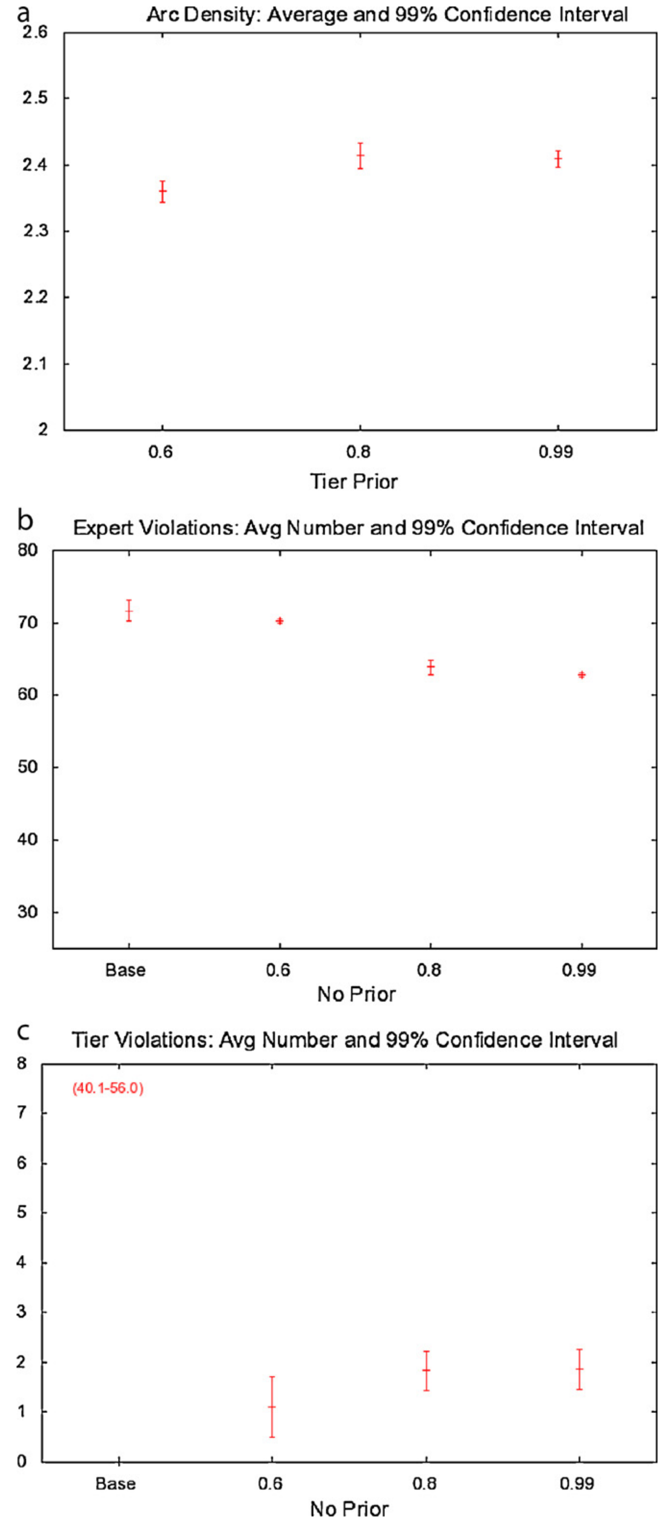


Fig. 14. Experiment 3 combined priors: varying confidence in expert pair-wise priors 0.6, 0.8, 0.99 (confidence in tier priors fixed at 0.9999) (a) average arc density, (b) average no. of violations of pair-wise priors, (c) average no. of tier prior violations (all with 99% confidence interval).

(80%) including ABLWALK → HADSTRKE, which violates the same tier prior and F (73.33%).

We next qualitatively examine the single BNs generated for threshold 70% for Experiments 1 and 3, which allows us to assess the benefit of incorporating the expert prior.

Table 3
Comparison of single BNs generated from Experiments 1 and 3: differences in arcs (arcs violating tiers indicated with *).

Pair	Tier–Tier	Expt 1 Single BN	Expt 3 Single BN
SITE -- SEX	bg -- bg		→
AGE -- ABLWALK	bg -- c		→
SEX -- HADSTRKE	bg -- p		→
HADHRTAT -- HADSTRKE	p -- p	→	
HADHRTAT -- HADDBTES	p -- p	←	
HADHRTAT -- ABLWALK	p -- c		←*
HADSTRKE -- ABLWALK	p -- c	→	←*
HADDBTES -- ABLWALK	p -- c		→
SHRBRLIE -- ABLWALK	p -- c	→	←
SHRBRLIE -- PRSSCHST	c -- c	→	←

First, we identified the differences between the two BNs, in terms of arc existence or direction, as shown in Table 3, and looked at the arcs in relation to tiers. The combined priors (Experiment 3) network had 3 more arcs in total (5 that were not in the Experiment 1 network, and without two that were in the Experiment 1 network), while the arc directions were also swapped for 3 arcs. There were only two tier violations in the combined single BN – HADHRTAT → ABLWALK and HADSTRKE → ABLWALK (both current → past), and none in the tiers only (Experiment 1) BN. Both the single BNs had the same 70 pairwise prior violations. The only arcs into the un-tiered nodes AGE, SMKGNW, BMIKind are background nodes; this suggests they might be considered part of the bg grouping. We also note that the arcs to EVRMARRD in the summary graph from Experiment 1 (as shown in Table 2) are not present in the Experiment 1 single BN, as their 0.57 frequency was less than the 70% threshold.

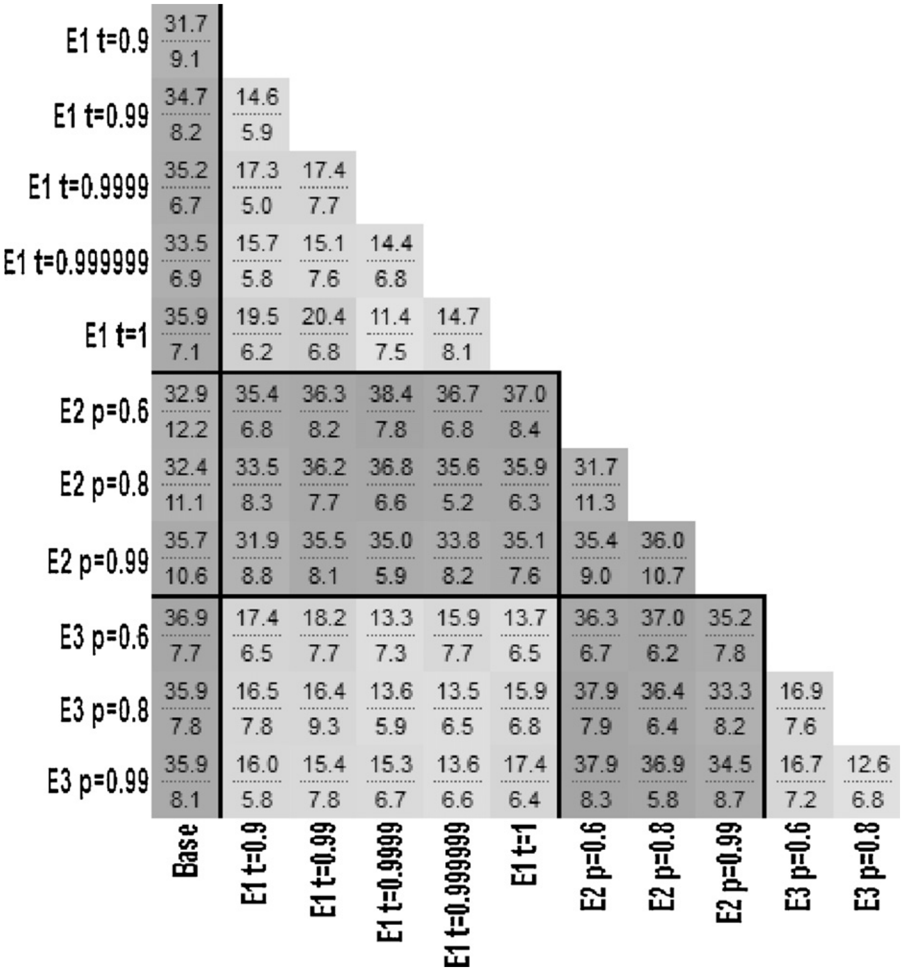
The breakdown of the arcs (i.e., across tiers or within tiers) is similar for the BNs from both experiments 1 and 3. There are relatively few arcs within tiers. There are also quite a few arcs from bg → c. This indicates that, although we provided the tiers bg ≺ p ≺ c, this does not generate a clear sequence of tiers; this makes sense given the definition of the tier relationship ≺ as forbidding a descendant relationship, but not enforcing an ancestor relationship. Finally, our expert qualitatively analysed the arcs which differed between the 2 single BNs, as follows.

7.1.1. *AGE → ABLWALK (Expt 3)*

This direct arc was a prior given by the expert. But when asked for further explanation, our expert noted that, recalling that that AGE is not the actual age of the subject, but rather the answer to the question *How old are you* (incorrect/correct or refused), getting one's age incorrect is a sign of dementia. Dementing patients are known to have diminished walking, but equally diminished walking is a known predictor of dementia. A possible mechanism for the direction learned by CaMML is "lack of volition"—you do not walk when you are dementing because you cannot be bothered. Overall, our expert agreed there should be an arc, but was neutral on the direction.

7.1.2. *HADHRTAT → HADSTRKE (Expt 1)*

Our expert saw no reason for there to be a direct arc, although these may be related by means of associated variables, some of which are measured in this model (high blood pressure, diabetes) and some of which are not measured in this model (cholesterol levels).



	Base	E1 t=0.9	E1 t=0.99	E1 t=0.9999	E1 t=0.999999	E1 t=1	E2 p=0.6	E2 p=0.8	E2 p=0.99	E3 p=0.6	E3 p=0.8
E1 t=0.9	31.7 9.1										
E1 t=0.99	34.7 8.2	14.6 5.9									
E1 t=0.9999	35.2 6.7	17.3 5.0	17.4 7.7								
E1 t=0.999999	33.5 6.9	15.7 5.8	15.1 7.6	14.4 6.8							
E1 t=1	35.9 7.1	19.5 6.2	20.4 6.8	11.4 7.5	14.7 8.1						
E2 p=0.6	32.9 12.2	35.4 6.8	36.3 8.2	38.4 7.8	36.7 6.8	37.0 8.4					
E2 p=0.8	32.4 11.1	33.5 8.3	36.2 7.7	36.8 6.6	35.6 5.2	35.9 6.3	31.7 11.3				
E2 p=0.99	35.7 10.6	31.9 8.8	35.5 8.1	35.0 5.9	33.8 8.2	35.1 7.6	35.4 9.0	36.0 10.7			
E3 p=0.6	36.9 7.7	17.4 6.5	18.2 7.7	13.3 7.3	15.9 7.7	13.7 6.5	36.3 6.7	37.0 6.2	35.2 7.8		
E3 p=0.8	35.9 7.8	16.5 7.8	16.4 9.3	13.6 5.9	13.5 6.5	15.9 6.8	37.9 7.9	36.4 6.4	33.3 8.2	16.9 7.6	
E3 p=0.99	35.9 8.1	16.0 5.8	15.4 7.8	15.3 6.7	13.6 6.6	17.4 6.4	37.9 8.3	36.9 5.8	34.5 8.7	16.7 7.2	12.6 6.8

Fig. 15. Edit distance grid: each cell contains ED (and standard deviation) between the learned BNs from those experiments.

7.1.3. *HADHRTAT → ABLWALK (Expt 3)*

This followed the expert's pairwise prior. Upon further consideration, our expert agreed that there are causal explanations for arcs in both directions: a sedentary life in which you cannot exercise because you cannot walk may lead to a heart attack (this is known mainly from associational studies, some randomised trials, e.g., MRFIT [52], did not find this), however having a heart attack might be associated with difficulty walking through heart failure, or through angina induced

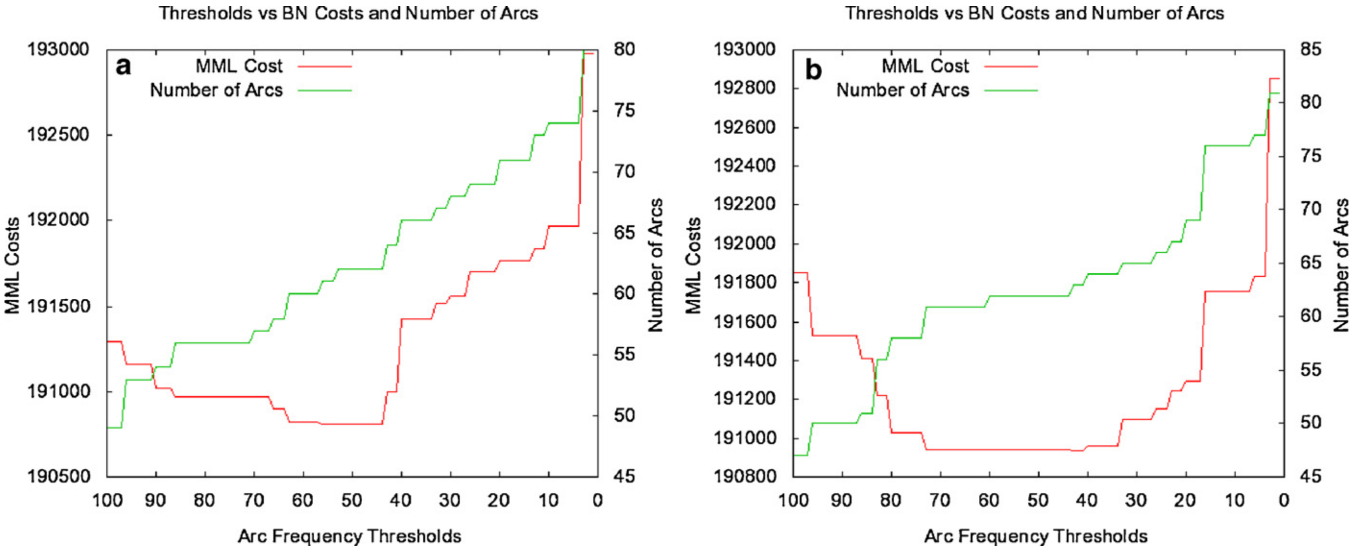


Fig. 16. Number of arcs and MML cost for single BNs generated for threshold frequency 0% to 100%, in 1% increments, from the summary matrix for (a) Experiment 1, tier prior conf = 0.9999 and (b) Experiment 3, tier prior conf = 0.9999 and pair-wise prior conf = 0.8.

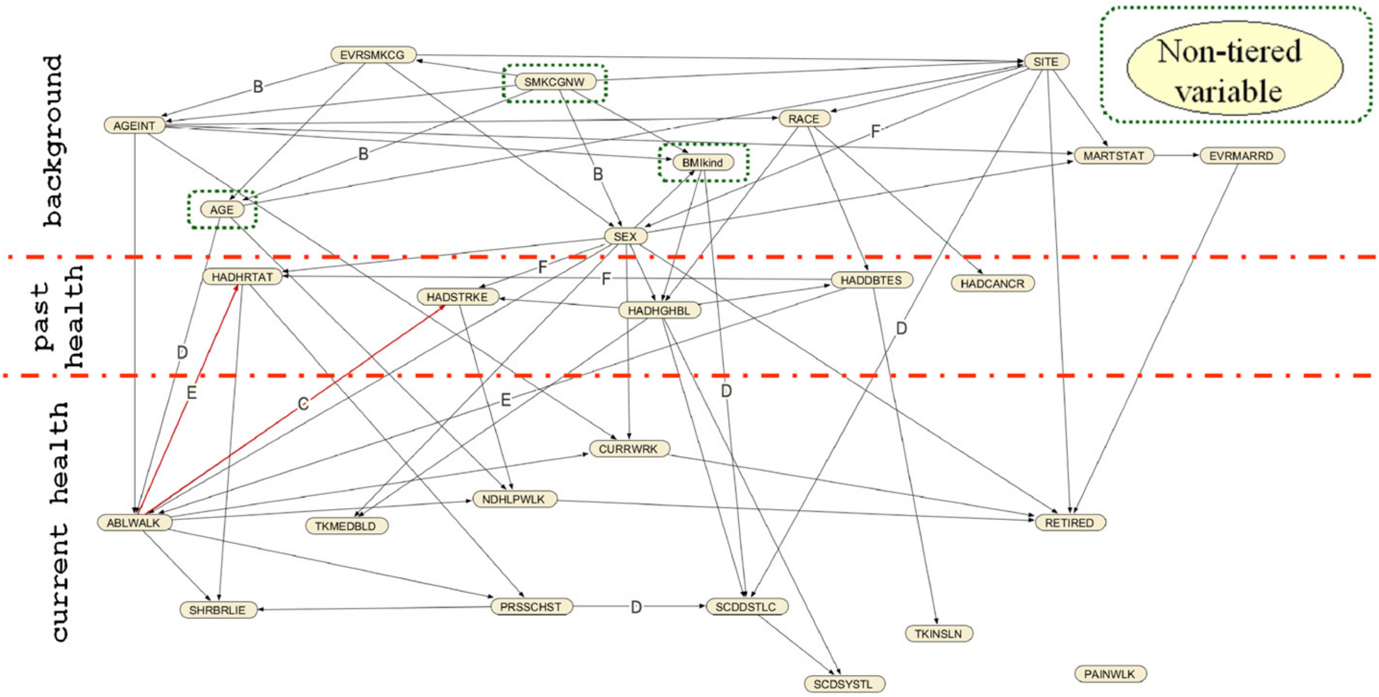


Fig. 17. Single BNs generated from the summary matrix for Experiment 3 (tier prior conf = 0.9999, pair-wise prior conf = 0.8), for thresholds 100% down to 70%. Tiers are separated by dotted lines, nodes covered with a dotted frame are non-tiered. Arc labels indicate threshold frequency for arc: B = 96.67%, C = 86.66%, D = 83.33%, E = 80%, F = 73.33%. Red arcs violate a structural prior.

by exercise, or through associated peripheral vascular disease.

7.1.4. HADSTRKE – ABLWALK

Our expert supports the HADSTRKE → ABLWALK arc (Experiment 1) rather than ABLWALK → HADSTRKE (Experiment 3), since there is often a direct causal mechanism of that kind. Our expert could not identify any direct causal relationship between stopping walking and having a stroke.

7.1.5. HADDBTES → ABLWALK (Expt 3)

Our expert was able to provide the following domain justifications for an arc between these, although was uncertain about what direction it should have (i.e., remaining with ∼ in original pairwise priors). First, increasing exercise (with walking the most common form of exercise) is known to be able to reduce the risk of diabetes (e.g., randomised trials in Scandinavia [53]). Hence not being able to walk half a mile (which is a relatively rigourous indicator of exercise tolerance) might be considered to be a potential risk for developing diabetes, i.e., supporting HADDBTES → ABLWALK.

However patients with diabetes have an increased risk for peripheral vascular disease with decreased circulation and decreased sensation problems in the legs,[10] which can lead to lower limb gangrene/infection, then toe or leg amputation—both of which clearly reduce the possibility of walking (i.e., a negative HADDBTES → ABLWALK relationship). In addition, progressive exercise is often advocated for patients with diabetes as it appears that it may slow the rate at which complications occur (including heart disease), which would be modelled as a positive HADDBTES → ABLWALK relationship. This signals the need for further analysis of this arc after parameterisation of the network.

Our expert found the lack of a HADDBTES – ABLWALK relationship in Experiment 1 surprising, as exercise is believed to be a moderately powerful controller of onset of diabetes (i.e., one of the reasons for an epidemic of type 2 diabetes is believed to be due to lack of exercise).

7.1.6. ABLWALK, PRSSCHST, SHRBRLIE

In both the Experiment 1 and Experiment 3 single BNs, these 3 nodes are connected to form a small subgraph, with ABLWALK → PRSSCHST in both. Our expert noted that PRSSCHST answered “Have you *ever* had pressure in your chest?”, which gives a lifetime of opportunities to have this symptom, whereas the other 2 variables refer to the present. This might suggest PRSSCHST should be concomitant or preceding the other. Our expert found no obvious direct causal explanations that would suggest one subnet is to be preferred to another. Clinically however it would usually be assumed that “pressure in the chest” indicated either angina, or, in the acute case, a heart attack. Following this, some degree of chronic heart failure signaled by SHRBRLIE is very possible, and the ability to walk 1/2 a mile would also definitely be compromised. Overall, the expert preferred the Experiment 3 structure, where ABLWALK is a parent to both PRSSCHST and SHRBRLIE, as the progress of heart failure is most likely to produce restriction of walking before shortness of breath lying down.

We note that in both subgraphs, all the nodes are *dependent*—adding evidence about any one variable will change the posteriors in the other two. Given our expert’s explanation of the clinical situation, we suggest that these nodes may in fact have a unobserved “common cause”, similar to Fig. 1(b). In such a structure, the child nodes are dependent, when the common cause is not known—which is exactly what the Experiment 1 and Experiment 3 single BNs show. (There may also be additional arcs between the children indicating a possible progression of symptoms.)

Overall, it is clear that the combined priors have produced only slight differences in the single BN structure, compared to

[10] This is a known association, however randomised trials have not been done!

tiers only. Although the differences where slight, in only one case, HADHRTAT → ABLWALK did the expert clearly prefer the tier only learned relationship. In all the others, the expert preferred the combined priors model, sometimes agreeing with both arc and direction, in other cases agreeing with the existence of a direct arc, even though unsure of best causal direction. Another outcome of this stage of the KEBN process were several changes to the expert's pairwise priors, which could be used for another iteration.

8. Conclusions and future work

In this paper we have presented a methodology for incorporating expert knowledge as structural priors when learning BNs. We have demonstrated its use in a medical case study of heart failure. Our methodology is an iterative one, reflecting the way in which BNs are developed in practice. It also makes explicit the possible interactions between expert elicitation and automated learning from data. We also presented novel visualisations of the learned networks, which support the interactive development process by allowing the knowledge engineers to assess intermediate results and revise experimental parameters. These visualisations could also assist comparisons of BN learning algorithms (e.g., [12]).

For problems with more than a small number of variables, the elicitation burden on experts of providing pairwise priors is heavy. This paper has shown how this burden can be reduced by a preliminary classification of the variables (e.g., as either background, past health or current health) which provides natural tier priors. These tiers are general enough to be a template for other medical BN applications. The experimental results show that, for our case study, the tier priors alone result in the majority of structural changes, compared to the base case learning from data only. The pairwise expert priors conflicted more with the data; however, when combined with the tier priors, they provided some structural differences that our expert felt better modelled the heart failure domain.

In this paper, as part of our comparative experimental study, we elicited and applied the pairwise priors separately, which allowed us to trial different confidence parameters. In practice, to reduce the elicitation burden, we suggest that tiers be elicited first and then used as constraints, with the expert only being asked about pairwise relationships across tiers and within tiers. We limited our study to priors elicited from a single expert; an obvious area for future investigation is how to use multiple experts both to spread the elicitation burden and to obtain 'consensus' priors of better quality.

We limited this case study to Stage 1 (structure building and evaluation) of the overall KEBN process. Further iterations within Stage 1 could use the pairwise priors revised in light of the expert's single BN evaluation (Section 7), or apply other evaluation methods (e.g., checking dependencies using the Matilda tool [48]). Of course, Stage 2 (parameter estimation and evaluation) must further be undertaken (possibly interleaved with more iterations within Stage 1) to produce a deployable BN. Here we have focused only on the model building task. There is also much work to be done before we see learned BNs embedded in a useful decision support tool (along the lines of TakeHeartII [54]) that will be adopted by physicians and used to improve their decision-making and clinical care.

Finally, we note that the usefulness of any learned BN is limited by the data available. We have seen with the Iowa dataset, which did not include variables on the actual disease, that parts of the learned structure do not reflect the understood medical causal process. To overcome this, we plan to extend CaMML to learn BNs with unobserved variables (e.g., [55]), including allowing experts to suggest these variables and provide structural priors for them.

Another area for future investigation is to extend CaMML to learn dynamic Bayesian networks, which have an explicit representation of time, not allowing arcs to go temporally backwards. This would explicitly provide temporal tiers that would never be violated.

Acknowledgements

This work has been partially funded by FEDER funds and the Spanish Government (MICINN) through project TIN2010-20900-C04-03 and UCLM under Project PL20091291. We thank Rodney O'Donnell for his assistance with CaMML, and the anonymous reviewers for their useful comments and suggestions.

References

[1] Pearl J. Probabilistic reasoning in intelligent systems. San Mateo, CA: Morgan Kaufmann; 1988.
[2] Korb KB, Nicholson AE. Bayesian artificial intelligence. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2010.
[3] Beinlich I, Suermondt H, Chavez R, Cooper G. The ALARM monitoring system. In: Hunter J, editor. Proceedings of the second European conference on artificial intelligence in medicine. Heildeberg: Springer-Verlag; 1992. p. 689–93.
[4] van der Gaag LC, Renooij S, Witteman CLM, Aleman BMP, Taal BG. Probabilities for a probabilistic network: A case-study in oesophageal cancer. Artificial Intelligence in Medicine 2002;25(2):123–48.
[5] Burnside BE, Rubin DL, Shachter R. A Bayesian network for mammography. In: Overhage JM, editor. Proceedings of the 2000 AMIA annual symposium. 2000. p. 106–10.
[6] Onisko A, Druzdzel M, Wasyluk H. A probabilistic model for diagnosis of liver disorders. In: Klopotek M, Michalewicz M, Ras Z, editors. Proceedings of the seventh symposium on intelligent information systems (IIS-98). 1998. p. 379–87.
[7] Twardy CR, Nicholson AE, Korb KB, McNeil J. Epidemiological data mining of cardiovascular Bayesian networks. Electronic Journal of Health Informatics 2006;1(1):1–13.
[8] Nicholson AE, Twardy CR, Korb KB, Hope LR. Decision support for clinical cardiovascular risk assessment, statistics in practice. John Wiley & Sons; 2008. p. 33–52, Ch 3.
[9] Wallace CS, Korb KB. Learning linear causal models by MML sampling. In: Gammerman A, editor. Causal Models and Intelligent Data Management. Heidelberg: Springer-Verlag; 1999. p. 89–111.
[10] Wallace, Korb, O'Donnell, Hope, Twardy. CaMML; 2005 http://www.datamining.monash.edu.au/software/camml (accessed: 4 February 2011).
[11] Knuiman MW, Vu HT, Bartholomew H. Multivariate risk estimation for coronary heart disease: the Busselton Health Study. Australian and New Zealand Journal of Public Health 1998;22:747–53.
[12] Acid S, de Campos LM, Fernàndez-Luna JM, Rodrìíguez S, Rodrìíguez JM, Salcedo JL. A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. Artificial Intelligence in Medicine 2004;30(3):215–32.
[13] Getoor L, Rhee JT, Koller D, Small P. Understanding tuberculosis epidemiology using structured statistical models. Artificial Intelligence in Medicine 2004;30(3):233–56.
[14] Promedas. http://www.promedas.nl (accessed: 4 February 2011).
[15] Hayduk LA. Equivalent models: TETRAD and model modification. Baltimore: Johns Hopkins University Press; 1996. p. 121–54, Ch. 4.
[16] O'Donnell R, Nicholson A, Han B, Korb K, Alam M, Hope L. Causal discovery with prior information. In: Sattar A, Kang BH, editors. AI 2006: advances in artificial intelligence (proceedings of the 19th Australian joint conference on advances in artificial intelligence [AI'06], Hobart, Australia, 4–8 December 2006), LNAI Series. Germany: Springer-Verlag; 2006. p. 1162–7.
[17] Antal P, Fannes G, Timmerman D, Moreau Y, De Moor B. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. Artificial Intelligence in Medicine 2004;30(3):257–81.
[18] Boneh T. Ontology and Bayesian decision networks for supporting the meteorological forecasting process. Ph.D. thesis. Clayton School of Information Technology, Monash University; 2010.
[19] Pollino C, Woodberry O, Nicholson A, Korb K, Hart BT. Parameterisation of a Bayesian network for use in an ecological risk management case study. Environmental Modelling and Software 2007;22(8):1140–52.
[20] Taylor JO, Wallace RB, Ostfeld AM, Blazer DG. Established populations for epidemiologic studies of the elderly, 1981–1993: [East Boston, Massachusetts, Iowa and Washington Counties, Iowa, New Haven, Connecticut, and North Central North Carolina]. Third ICPSR version, Interuniversity Consortium for Political and Social Research (http://dx.doi.org/10.3886/ICPSR09915).
[21] McDonagh T, Morrison C, Lawrence A, Ford I, Tunstall-Pedoe H, Mortan J, et al. Symptomatic and asymptomatic left ventricular dysfunction in an urban population. The Lancet 1997;350:829–33.
[22] Davie A, Francis C, Caruana L, Sutherland G, McMurray JJ. Assessing diagnosis in heart failure: which features are any use? QJM 1997;90(5):335–9.
[23] Jessup M, Brozena S. Heart failure. New England Journal of Medicine 2003;348:2007–18.

[24] Lauritzen SL, Spiegelhalter D. Local computation with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) 1988;50(2):157–224.

[25] Shwe M, Middleton B, Heckerman D, Henrion E, Horvitz M, Lehmann H, et al. Probabilistic diagnosis using a reformulation of the INTERNIS T-1/QMR knowledge base I. The probabilistic model and inference algorithms. Methods in Information in Medicine 1991;30:241–55.

[26] Pradham M, Provan G, Middleton B, Henrion M. Knowledge engineering for large belief networks. In: de Mantaras L, Poole D, editors. UAI94—proceedings of the 10th conference on uncertainty in artificial intelligence. 1994. p. 484–90.

[27] Jensen FV, Lauritzen SL, Olesen KG. Bayesian updating in causal probabilistic networks by local computations. Computational Statistics Quarterly 1990;4:269–82.

[28] Chickering DM. A tranformational characterization of equivalent Bayesian network structures. In: UAI95—proceedings of the 11th conference on uncertainty in artificial intelligence. 1995. p. 87–98.

[29] Spirtes P, Glymour C, Scheines R. Causation, prediction and search. 2nd ed. MIT Press; 2000.

[30] Yehezkel R, Lerner B. Bayesian network structure learning by recursive autonomy identification. Journal of Machine Learning Research 2009;10: 1527–70.

[31] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 1992;9:309–47.

[32] Wallace C, Korb K. Learning linear causal models by MML sampling. In: Gammerman A, editor. Causal models and intelligent data management. Springer; 1999. p. 89–111.

[33] Chickering DM. Optimal structure identification with greedy search. Journal of Machine Learning Research 2003;3:507–54.

[34] Wallace CS, Boulton DM. An information measure for classification. The Computer Journal 1968;11:185–94.

[35] Wallace CS. Statistical and inductive inference by minimum message length. Berlin, Germany: Springer; 2005.

[36] Rissanen J. Modeling by shortest data description. Automatica 1978;14:465–71.

[37] Lam W, Bacchus F. Learning Bayesian belief networks: an approach based on the MDL principle. Computational Intelligence 1993;10:269–93.

[38] Suzuki J. Learning Bayesian belief networks based on the Minimum Description Length principle. In: ICML-96—proceedings of the 13th international conference on machine learning. 1996. p. 462–70.

[39] Neil JR, Korb KB. The evolution of causal models: a comparison of Bayesian metrics and structure priors. In: Zhong N, Zhous L, editors. Methodologies for knowledge discovery and data mining: third Pacific-Asia conference. Heidelberg: Springer Verlag; 1999. p. 432–7.

[40] Dai H, Korb KB, Wallace CS, Wu X. A study of casual discovery with weak links and small samples. In: IJCAI97—proceedings of the fifteenth international joint conference on artificial intelligence. 1997. p. 1304–9.

[41] Neil JR, Wallace CS, Korb KB. Learning Bayesian networks with restricted causal interactions. In: Laskey, Prade, editors. UAI99—proceedings of the fifteenth conference on uncertainty in artificial intelligence. 1999. p. 486–93.

[42] O'Donnell R. Learning flexible causal models by MML. Ph.D. thesis. Clayton School of Information Technology, Monash University; 2010.

[43] Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning 1995;20: 197–243.

[44] HUGIN EXPERT A/S, http://www.hugin.com (accessed: 4 February 2011).

[45] Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. In: de Mantras L, Poole D, editors. Proceedings of the tenth conference on uncertainty in artificial intelligence. 1994. p. 293–301.

[46] Castelo R, Siebes A. Priors on network structures. Interational Journal of Approximate Reasoning 2000;24(1):39–57.

[47] Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Morgan Kaufmann: Series in data management systems; 2006.

[48] Boneh T, Nicholson AE, Sonenberg EA. Matilda: a visual tool for modeling with Bayesian networks. International Journal of Intelligent Systems 2006;21(11):1127–50.

[49] Linstone HA, Turoff M. The Delphi method: techniques and applications. Reading, MA: Adison-Wesley; 1975.

[50] Hope L, Korb K. A Bayesian metric for evaluating machine learning algorithms. In: Webb G, Yu X, editors. Proc. of the 17th Australian joint conf. on advances in art. intelligence (AI'04), vol. 3229 of LNCS/LNAI Series. Berlin, Germany: Springer-Verlag; 2004. p. 991–7.

[51] Hodges AP, Dai D, Xiang Z, Woolf P, Xi C, He Y. Bayesian network expansion identifies new ros and biofilm regulators. PLoS ONE 2010;5(3):e9513.

[52] Kannel WB, Neaton JD, Wentworth D, Thomas HE, Stamler J, Hulley SB. Overall and coronary heart disease mortality rates in relation to major risk factors in 325, 348 men screend for the MRFIT. American Heart Journal 1986;112:825–36.

[53] Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Pirjo I, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. The New England Journal of Medicine 2001;344(18):1343–50.

[54] Nicholson AE, Twardy CR, Korb KB, Hope LR. Decision support for clinical cardiovascular risk assessment. In: Pourret O, Naim P, Marcot B, editors. Bayesian networks: a practical guide to applications, statistics in practice. Wiley; 2008. p. 33–52.

[55] Zhang NL, Nielsen TD, Jensen FV. Latent variable discovery in classification models. Artificial Intelligence in Medicine 2004;30(3):283–99.