Using Temporal Context-Specific Independence Information in the Exploratory Analysis of Disease Processes

Stefan Visscher[1], Peter Lucas[2], Ildikó Flesch[2], and Karin Schurink[1]

[1] Department of Internal Medicine and Infectious Diseases, University Medical Center Utrecht, The Netherlands
{S.Visscher,K.Schurink}@umcutrecht.nl
[2] Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands
{peterl,ildiko}@cs.ru.nl

Abstract. Disease processes in patients are temporal in nature and involve uncertainty. It is necessary to gain insight into these processes when aiming at improving the diagnosis, treatment and prognosis of disease in patients. One way to achieve these aims is by explicitly modelling disease processes; several researchers have advocated the use of dynamic Bayesian networks for this purpose because of the versatility and expressiveness of this time-oriented probabilistic formalism. In the research described in this paper, we investigate the role of context-specific independence information in modelling the evolution of disease. The hypothesis tested was that within similar populations of patients differences in the learnt structure of a dynamic Bayesian network may result, depending on whether or not patients have a particular disease. This is an example of temporal context-specific independence information. We have tested and confirmed this hypothesis using a constraint-based Bayesian network structure learning algorithm which supports incorporating background knowledge into the learning process. Clinical data of mechanically-ventilated ICU patients, some of whom developed ventilator-associated pneumonia, were used for that purpose.

1 Introduction

Bayesian networks are known to yield representations that are well suited as a basis for medical decision making [1]. Reasoning with a Bayesian network, which is done by filling in data of a patient into the network and computing posterior probability distributions, often yields considerable insight into the disease process of a patient, as well as concerning the way the disease process can be influenced by the selection of appropriate treatment. However, knowledge of the temporal nature of a disease process may also be relevant in this respect, in which case *dynamic* Bayesian networks are often selected for the construction of models. Such models can be used for temporal reasoning in clinical decision-support systems, as the formalism takes into account the notion of time [2,3,4,5]. Bayesian networks that ignore the notion of time are called *static*.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 87–96, 2007.

So far, Bayesian networks have in particular been popular as models for uncertainty reasoning in clinical decision-support systems; they have been less popular as tools for the analysis of clinical data, despite the availability of a wide range of Bayesian network structure and parameter learning techniques [6]. This is somewhat surprising as the statistical nature of Bayesian networks would render them in principle as useful as data-analytical tools as, say, logistic regression, one of the main statistical tools of multivariate clinical data analysis.

We believe that the reasons why Bayesian networks, both static and dynamic, are used so rarely for data analysis in medicine are threefold: (1) in particular static Bayesian networks are difficult to interpret, as the direction of the arcs is often counterintuitive; (2) whereas dynamic Bayesian networks have the advantage that the direction of some of the arcs is in accordance to the order of time, their structure is usually restricted to being *repetitive* [7], which may not be compatible with the clinical problem at hand; (3) the conditional independences modelled in Bayesian networks only concern random variables and not their individual values; however, in medicine it is often the *context*, i.e., the specific values random variables take, that determine how things relate to each other. Context-specific independences and dependences can be modelled by extensions to Bayesian networks, such as by *multinet* models [8].

In this paper, we demonstrate that by using non-repetitive dynamic Bayesian multi-networks in conjunction with context-specific independence information, an analytical tool results that does indeed yield insight into the evolution of a disease, in comparison to other diseases in a related population of patients. We exploit a constraint-based learning algorithm for that purpose, as these structure learning algorithms allow for the easy incorporation of medical background knowledge in the learning process. The ideas are illustrated by the analysis of temporal data of patients in the Intensive Care Unit (ICU), who either have developed ventilator-associated pneumonia, or VAP for short, and ICU patients without VAP. As only some 10-15% of ICU patients will develop VAP, it was also necessary to exploit background knowledge in the learning process, as sometimes clinically obvious relationships cannot be learnt from the data due to the sparsity of data for a particular type of patient.

The paper is organised as follows. In Section 2, Bayesian networks, dynamic Bayesian networks and context-specific independence are briefly reviewed. Next, in Section 3, the basic theory underlying constraint-based structure learning is reviewed. Finally, in Section 4 we discuss the results achieved. The paper is rounded-off with some conclusions in Section 5.

2 Preliminaries

We briefly review the theory dynamic Bayesian networks, as discussed in more detail in [7]. Furthermore, the medical domain of ventilator-associated pneumonia, is described.

2.1 Dynamic Bayesian Networks

A *Bayesian network* $\mathcal{B} = (G, P)$, BN for short, is a joint probability distribution P of a set of random variables X with an associated acyclic directed graph $G = (V, A)$, where P is assumed to be decomposed into a set of conditional probability distributions in accordance to the structure of G. The random variables X and the vertices V have a 1–1 correspondence; thus we sometimes write X_W, $W \subseteq V$, for the random variables corresponding to the vertices W. Finally, $dom(X)$ denotes the domain of the set of random variables X (a Cartesian product).

Dynamic Bayesian networks (DBNs) are an extension of ordinary Bayesian networks and allow for modelling uncertainty involved in processes regarding the dimension of time. Usually, a DBN is described in terms of a timeslice that has a fixed structure and is repeated several times, i.e., the DBN has a *repetitive* structure [9]. We, however, are convinced that disease processes are more complicated than that in the sense that independences may change over time and, therefore, a repetitive DBN would not suffice in every domain. This motivated some of us to develop a theory of modularisation of DBNs, with both repetitive and non-repetitive DBNs as special cases [7]. Evidence of the practical usefulness of non-repetitive DBNs has also come from work by Tucker et al. [10].

For the formal representation of the uncertain relations between variables over time, we need the following notions. Let T denote the (discrete and finite) time axis. Independence relationships between random variables with the *same* time point t are represented by means of an acyclic directed graph (ADG) $G_t = (V_t, A_t^a)$, called a *timeslice*, with V_t denoting a set of vertices and $A_t^a \subseteq V_t \times V_t$ a set of *atemporal arcs*. Between timeslices, vertices corresponding to random variables may be linked to each other by means of so-called *temporal arcs*. Thus, a DBN consists of two parts: (1) an atemporal part (the timeslices), and (2) a temporal part. First, we consider the atemporal part.

Definition 1. ***(timeslice and atemporal arcs)*** *An ADG $G_t = (V_t, A_t^a)$, with the set of vertices V_t and the set of* atemporal arcs *$A_t^a \subseteq V_t \times V_t$, $t \in T$, is called a* timeslice *at time point t.*

The set of all timeslices G of a DBN is taken as:

$$G = \{G_t \mid t \in T\} = \{(V_t, A_t^a) \mid t \in T\} = (V_T, A_T^a). \tag{1}$$

Let G_t and $G_{t'}$, $t, t' \in T$, be two timeslices. Then, an arc $(u_t, v_{t'})$ with $t < t'$ is called a *temporal arc*. The set of temporal arcs of an ADG is denoted by A^t. Thus, temporal arcs connect timeslices with strict direction from the past to the future.

Definition 2. ***(temporal network)*** *A* temporal network *N is defined as a pair $N = (V_T, A)$, where $G = (V_T, A_T^a)$ and $A = A_T^a \cup A^t$, with A_T^a denoting the set of timeslices.*

Clearly, a temporal network N is also an ADG. A *dynamic Bayesian network* (DBN) is now defined as a pair $\mathcal{DBN} = (N, P)$, where P is the joint probability distribution (JPD) on X_{V_T}.

2.2 Context-Specific Independences

Two sets of random variables X and Y are said to be *conditionally independent* given a third set of random variables Z, denoted by $X \perp\!\!\!\perp_P Y \mid Z$, if it holds that

$$P(X \mid Y, Z) = P(X \mid Z)$$

if $P(Y, Z) > 0$. Such conditional independence statements cannot only be represented in the form of probability distributions P; they can also be read-off from the graphical structure of an associated ADG G using the notion of d-separation. Then, two disjoint sets of vertices A and B in G are said to be *d-separated* given a third disjoint set of vertices C, denoted by $A \perp\!\!\!\perp_G B \mid C$, if each (undirected) path from a vertex in A to a vertex in B is blocked by a vertex in C, taking into account paths with so-called v-structures (i.e., subgraphs of the form $\rightarrow \cdot \leftarrow$).

For Bayesian networks $\mathcal{B} = (G, P)$, it holds that if $A \perp\!\!\!\perp_G B \mid C$ holds, then $X_A \perp\!\!\!\perp_P X_B \mid X_C$ should also be satisfied. It is said that G is an *independence map* of P. Similar, temporal and atemporal, notions of d-separation have been developed for dynamic Bayesian networks, where the *atemporal d-separation* relationship $\perp\!\!\!\perp_G$ is defined for the part of the dynamic Bayesian network where the temporal arcs are ignored, and *temporal d-separation*, denoted by $\perp\!\!\!\perp_{N_{|\Theta}}$, is defined by always taking into account at least one temporal arc when investigating blockage (for details, cf. [7]). Clearly, atemporal d-separation $\perp\!\!\!\perp_G$ can be defined in terms of atemporal d-separation for individual timeslices, i.e., in terms of $\perp\!\!\!\perp_{G_t}$, $t \in T$.

Despite the fact that temporal and atemporal notions of d-separation allow for the study of interesting independence patterns in dynamic Bayesian networks, we believe that many of these patterns are context specific, i.e., independence information may change for particular values of random variables. Formally, a set of variables Y is *conditionally context-specific independent* of a set of variables W given a third set Z in the context φ, written $Y \perp\!\!\!\perp_P W \mid Z; \varphi$, where φ is a nonempty set of random variables U with values u, i.e., $\varphi \equiv U = u$, if $P(Y \mid W, Z, \varphi) = P(Y \mid Z, \varphi)$ and $P(Y \mid W, Z, \varphi') \neq P(Y \mid Z, \varphi')$ for $\varphi' \equiv U = u'$, $u' \neq u$ [11]. For discrete random variables X with finite domain, it is possible to associate an ADG G^{φ} with every context φ. The result is called a *Bayesian multinetwork* $\mathcal{B} = (G, P)$ with $G = \{G^{\varphi} \mid \varphi \equiv X = x, x \in dom(X)\}$. Dynamic Bayesian multinetworks can be be defined along similar lines.

2.3 Ventilator-Associated Pneumonia

Ventilator-associated pneumonia (VAP) occurs in mechanically-ventilated ICU patients. Clinical symptoms, such as fever, indicating that this bacterial infection is present or developing, are usually not very specific. Important symptoms and signs, providing evidence for the development of VAP, include *body temperature*, amount and colour of *sputum*, radiological *signs* on the chest X-ray, duration of *mechanical ventilation*, number of *leukocytes* [12], and abnormal ratio between the arterial oxygen pressure and the fractional inspired oxygen level

(pO_2/FiO_2-ratio). Some of these signs and symptoms, such as fever and number of leukocytes, are due to the fact that VAP is an infectious disease, whereas others, such as increased amount of sputum, abnormal chest X-ray and changed pO_2/FiO_2-ratio are due to the pulmonary location of the infection.

3 Constraint-Based Structure Learning

As only 10-15% of the ICU patients develop VAP, it was unlikely that we would have been able to collect sufficient amount of data for patient with VAP, despite the fact that the amount of data collected for the entire ICU population was large. For situations where data are sparse, it is normally difficult to learn independence relations from the data. However, lack of data can be compensated, in principle, by augmenting the learning process through the exploitation of background knowledge. This is exactly what we have done. Learning algorithms that allow easy incorporation of background knowledge into the learning process are called *constraint based*. These algorithms derive a set of conditional independence statements from the data, taking supplied dependence and independence information as additional constraints, and build a structure with d-separation properties corresponding to the independence information available.

3.1 The NPC Algorithm

One of the best constraint-based Bayesian network structure learning algorithms available is the NPC algorithm. NPC stands for 'Necessary Path Condition'; it is a criterion that has been added to an earlier constraint-based algorithm, PC, by researchers at Siemens in Munich [13]. The algorithm is a variant of the CI algorithm by Verma and Pearl [14], and works as follows:

1. **Automatic phase:**
   (a) An undirected graph $H = (V, E)$, called *skeleton*, is derived through computation of the score $g_{\chi^2,\alpha}(X, Y, S)$, for pairs X, Y of random variables and the set of random variables S (with $X, Y \notin S$). The function $g_{\chi^2,\alpha}$ is based on the χ^2 test with significance level α [13]. Typically, one takes $\alpha \leq 0.05$. If $g_{\chi^2,\alpha}(X, Y, S) > 0$ then the conditional independence hypothesis $X \perp\!\!\!\perp_P Y \mid S$ is rejected.
   (b) Modify subgraphs $X - Y - Z$ of H into $X \rightarrow Y \leftarrow Z$, if $X - Z \notin E$ and $X \not\perp\!\!\!\perp_P Z \mid S$, with $Y \in S$, using the same scoring function $g_{\chi^2,\alpha}$.
   (c) Orientate the remaining lines as to obtain arcs, where the creation of cycles in the resulting directed graph is avoided.
2. **User interaction phase:** To resolve inconsistencies in the conditional (in)dependence statements, the NPC algorithm, unlike the PC algorithm where in case of uncertain dependences directionality is chosen randomly, relies on user interaction where the user gets the opportunity to decide on the addition, removal and orientation of arcs. In addition, α can be arbitrarily chosen, so that lines with calculated p-value (called p below) larger than α are excluded.

In our domain, the direction of an arc has been determined by the use of background knowledge. By doing so, cause and effect can be distinguished. For example, when the NPC algorithm indicates a relation between 'VAP' and 'temperature', the most logical order is that 'VAP' should be parent of 'temperature', as when a patient suffers from VAP, normally, the temperature increases due to fever, and not the other way round. Also, it is possible to supply known relations at the start of the NPC learning process. By doing so, we were able to include constraints that have already been proved to exist and have been described in literature. We used the implementation of the NPC algorithm available in the Hugin tool set [15] for our research. This includes the EM algorithm for parameter learning.

3.2 Data

A dataset D with temporal data of ICU patients, containing 17710 records, was used. Each record represents data of one patient in the ICU during a period of 24 hours. The database contains 2424 admissions to the ICU. For 157 of these patient episodes, VAP was diagnosed by two infectious-disease specialists. From dataset D three subsets were created: D_{vap} containing data of all 157 VAP patients; $D_{\overline{\text{vap}}}$ containing data of patients who were not diagnosed with VAP (so-called controls). Each patient with VAP was matched to 3 control patients, with a similar duration of mechanical ventilation on the days of matching. D_{VAP} contained data of both VAP and control patients, i.e., $D_{\text{VAP}} = D_{\text{vap}} \cup D_{\overline{\text{vap}}}$. Each dataset contained data of 4 consecutive days, each representing a timeslice: t_0 was either the day on which VAP was diagnosed or the day of matching and t_{-3}, t_{-2}, t_{-1} were the three days preceding t_0.

3.3 Procedure of DBN Construction

The construction of the context-specific and combined DBNs was performed as follows:

1. Using the NPC algorithm, under the first author's supervision, the atemporal arcs between vertices in each separate timeslice were determined.
2. In the next run of the NPC algorithm, the temporal relationships of all variables were explored, taking into account the structure of the timeslices.
3. Medical background knowledge was sometimes employed to decide about the direction of arcs, or inclusion or deletion of arcs. However, this was only employed when the algorithm was unable to decide about the inclusion and direction of an arc. The direction of arcs was decided on by looking at the network in terms of cause–effect relationships. Expertise of a medical infectious disease specialist was used for that purpose.

4 Results

Based on the three databases, D_{vap}, $D_{\overline{\text{vap}}}$ and D_{VAP}, three DBNs were constructed using the NPC learning algorithm. As described above, atemporal subgraphs were obtained separately, and then combined to a DBN by learning temporal arcs

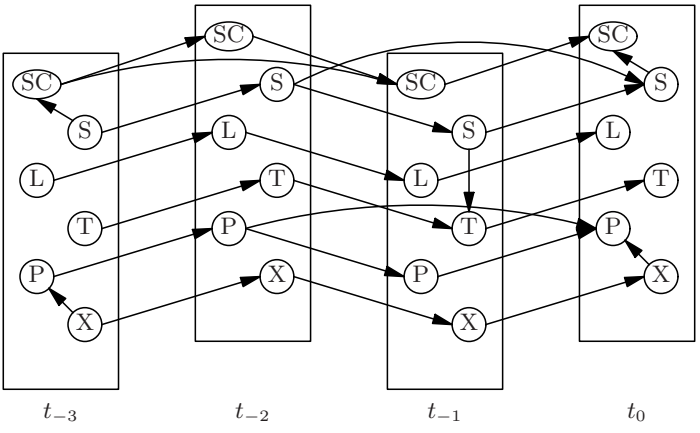

Fig. 1. G_{vap} : Independences obtained for VAP patients. Abbreviations: SC: sputum colour; S: sputum; L: leukocytosis; T: temperature; P: pO_2/FiO_2; X: chest X-ray.

(taking into account the known timeslice structures). Sometimes uncertain arcs were removed (user interaction phase). For example, an arc between the variable *temperature* in timeslice t_0 and variable *leukocytosis* in timeslice t_{-2}, did not seem clinically relevant and was, therefore, excluded.

4.1 VAP Patients (D_{vap})

The timeslices (atemporal subgraphs) for the four different time points show different independences. For example, for t_{-3} an arc between *sputum* and *sputum-colour* ($p = 0.05$) was suggested by the NPC algorithm, whereas for t_{-1} and for t_{-2} that same relation was absent, but for t_0 it was again present. Also, an arc ($p = 0.02$) between *chest X-ray* and pO_2/FiO_2 (as explained, a measurement of the lungs' functions) was often found, as well as between *temperature* and *sputum-colour* ($p = 0.05$). As the last arc was not considered clinically relevant, it was excluded from the models. All temporal arcs proved to have high significance ($p < 10^{-7}$). Combining the atemporal en temporal parts resulted in a DBN, called G_{vap}, shown in Fig. 1, that includes all signs and symptoms describing the course of the development of VAP.

4.2 Patients Not Diagnosed with VAP ($D_{\overline{\text{vap}}}$)

The timeslices for the non-VAP patients were similar, but not identical, to those for VAP patients; in particular, arcs between *chest X-ray* and pO_2/FiO_2, and between *sputum* and *sputum-colour* were found. The only difference was that the strength of the arcs increased towards the time point matching the day of VAP, i.e., t_0, that is, $p(t_{-3}) \approx 10^{-2}$, $p(t_{-2}) \approx 10^{-3}$, $p(t_{-1}) \approx 10^{-4}$ and $p(t_0) \approx 10^{-5}$. Thus, $p(t_{-3}) > p(t_{-2}) > p(t_{-1}) > p(t_0)$. Temporal arcs were suggested between timeslices t_{-2} and t_{-1} for the variables *chest X-ray*, *sputum* and pO_2/FiO_2

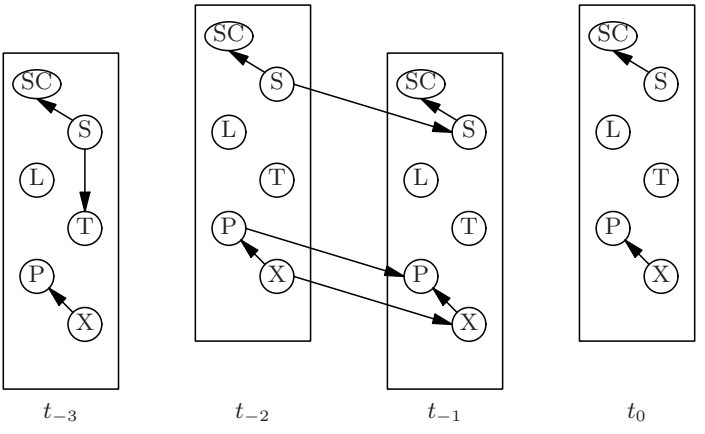

Fig. 2. $G_{\overline{vap}}$: Independences obtained for patients *not* diagnosed with VAP. Abbreviations: SC: sputum colour; S: sputum; L: leukocytosis; T: temperature; P: pO_2/FiO_2; X: chest X-ray.

only. Moreover, these temporal relations proved to be less strong, i.e., $p \approx 0.01$, compared to the temporal relations in the context of VAP. Combining both temporal and atemporal structures resulted in DBN called $G_{\overline{vap}}$, shown in Fig. 2 with again, $\overline{vap}$ representing the matched control patients.

4.3 Patients With and Without VAP ($D_{\mathbf{VAP}}$)

As a model only suitable for VAP patients would not be useful in practice, combining the two datasets mentioned above for building a DBN for VAP and

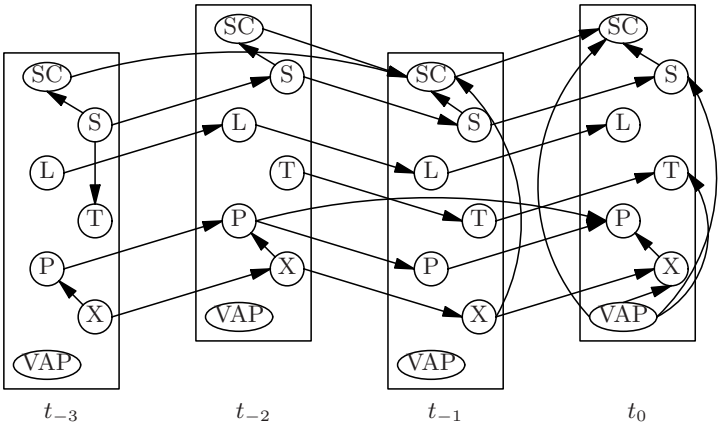

Fig. 3. G_{VAP} : Independence model for variables (excluding context-specific independences). Abbreviations: SC: sputum colour; S: sputum; L: leukocytosis; T: temperature; P: pO_2/FiO_2; X: chest X-ray; VAP: ventilator-associated pneumonia.

non-VAP at the same time yields yet another view on structure learning. Structure learning based on the database including data of VAP as well as non-VAP patients resulted in a combination of the two DBNs G_{vap} and $G_{\overline{vap}}$, from here denoted by G_{VAP}. The temporal arcs were almost identical to those of G_{vap}, though less strong ($p \approx 10^{-4}$). The atemporal arcs had strong correlations and were, not surprisingly, found between the variables *chest X-ray* and pO_2/FiO_2 ($p \approx 10^{-3}$) and between *sputum* and *sputum colour* ($p \approx 10^{-3}$). In all, the temporal arcs again proved to be stronger than the atemporal arcs. The resulting model is shown in Fig. 3. This DBN clearly shows that much of the clarity of the original context-specific DBNs was lost, and that it is no longer possible to gain insight into the development of VAP and non-VAP separately.

5 Conclusions and Discussion

The hypothesis underlying the research described in this paper was that in order to obtain insight into the evolution of disease processes, it is not merely necessary to explicitly model time, but also to consider context-specific independence information. The results we have obtained confirm this hypothesis, and to the best of our knowledge, this is the first paper combining context-specific independence and dynamic Bayesian networks.

The NPC learning algorithm proved to be useful, as it allowed for the incorporation of background knowledge, without which it would have been difficult to obtain clinically meaningful results. This algorithm combines the virtues of offering the capability of automatic learning of independence information from data, whereas uncertainty regarding both the presence of dependences and the directionality of arcs can be resolved by the user. Thus, the algorithm offers a natural role for the incorporation of expert background knowledge in the learning process.

The results obtained for the ICU domain show that signs and symptoms of patients known to develop VAP proved to have strong temporal relationships, whereas the temporal relationships between the signs and symptoms of patients not diagnosed with VAP were very weak. The combined model G_{VAP} included independences from both the VAP and non-VAP models. However, the model was not merely a union of the two independence sets.

When comparing the atemporal parts of the networks, change in the independence information as a function of time was observed. This strengthens our belief that when constructing a dynamic Bayesian model, it is not sufficient to resort to repetitive DBNs if one wishes to obtain models that capture the characteristics of the domain; rather, non-repetitive DBNs should be investigated as well [7]. Further research is needed to evaluate our findings regarding the temporal behaviour of both models for VAP patients and non-VAP patients. It would, for example, be interesting to investigate the predictive value of an increasing body temperature of an ICU patient in relationship to the development of VAP. Moreover, a more detailed comparison of the ADGs of VAP and non-VAP for larger datasets may give more insight into the course of the disease process of VAP.

In conclusion, the combination of a general theory of DBNs, where repetitive and non-repetitive DBNs are both special cases, with the exploitation of context-specific independence information proved to yield a powerful data-analysis tool.

Acknowledgements

This study was supported by a grant from NWO (project number 634.000.026).

References

1. Lucas, P.J.F., van der Gaag, L.C., Abu-Hanna, A.: Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine 30(3), 201–214 (2004)
2. Augusto, J.C.: Temporal reasoning for decision support in medicine. Artificial Intelligence in Medicine 33(1), 1–24 (2005)
3. Dagum, P., Galper, A.: Forecasting sleep apnea with dynamic models. In: Proceedings of UAI-1994, pp. 64–71 (1994)
4. Adlassnig, K.-P., Combi, C., Das, A.K., Keravnou, E.T., Pozzi, G.: Temporal representation and reasoning in medicine: Research directions and challenges. Artificial Intelligence in Medicine 38(2), 101–113 (2006)
5. Provan, G., Clarke, J.R.: Dynamic network construction and updating techniques for the diagnosis of acute abdominal pain. PAMI 15(3), 299–307 (1993)
6. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proceedings of the 14th UAI, pp. 139–147 (1998)
7. Flesch, I., Lucas, P.J.F., Visscher, S.: On the modularisation of independence in dynamic Bayesian networks. In: Proceedings of BNAIC-2006, pp. 133–140 (2006)
8. Geiger, D., Heckerman, D.: Knowledge representation and inference in similarity networks and Bayesian multinets. Artificial Intelligence 82, 45–74 (1996)
9. Labatut, V., Pastor, J., Ruff, S., Démonet, J.F., Celsis, P.: Cerebral modeling and dynamic Bayesian networks. Artificial Intelligence in Medicine 39, 119–139 (2004)
10. Tucker, A., Liu, X.: Learning dynamic Bayesian networks from multivariate time series with changing dependencies. In: Proceedings of the 5th International Symposium on Intelligent Data Analysis, Berlin, Germany (2003)
11. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: Proceedings of UAI-1996, pp. 64–72 (1996)
12. Bonten, M.J.M., Kollef, M.H., Hall, J.B.: Risk factors for Ventilator-Associated Pneumonia: from epidemiology to patient management. Clinical Infectious Diseases 38(8) (2004)
13. Steck, H.: Constraint-Based Structural Learning in Bayesian Networks using Finite Data Sets. PhD thesis, Institut für Informatik der Technischen Universität München (1999)
14. Verma, T.S., Pearl, J.: Equivalence synthesis of causal models. In: Proceedings of UAI-1990, pp. 220–227. Morgan Kaufmann, San Francisco (1990)
15. Madsen, A.L., Lang, M., Kjærulff, U.B., Jensen, F.: The hugin tool for learning Bayesian networks. In: Proceedings of 7th ECSQARU, pp. 594–605 (2003), URL: `http://www.hugin.com`