Induction of Partial Orders to Predict Patient Evolutions in Medicine

John A. Bohada, David Riaño, and Francis Real

Research Group on Artificial Intelligence
Departament of Computer Sciences and Mathematics, Rovira i Virgili University
Av. Països Catalans 26, 43007 Tarragona, Spain
{john.bohada,david.riano,francis.real}@urv.net

Abstract. In medicine, prognosis is the task of predicting the probable course and outcome of a disease. Questions like, is a patient going to improve?, what is his/her chance of recovery?, and how likely a relapse is? are common and they rely on the concept of state. The feasible states of a disease define a partial order structure with extreme states those of 'cure' and 'death'; improving, recovering, and survival meaning particular transitions between states of the partial order. In spite of this, it is not usual in medicine to find an explicit representation either of the states or of the states partial order for many diseases. On the contrary, the variables (e.g. signs and symptoms) related to a disease and their normality and abnormality values are broadly agreed. Here, an inductive algorithm is introduced that generates partial orders from a data matrix containing information about the patient-professional encounters, and the normality functions of each one of these disease variables.

1 Introduction

In medicine, prognosis is the process by which the probable course and outcome of a disease is predicted. Statistics and Artificial Intelligence have traditionally faced this process with several *methodologies* as survival analysis, logistic regression, Bayesian Networks, Artificial Neural Networks, Genetic Algorithms, and Decision Trees as [4] and [3] report. All these methodologies have been applied to predict *medical facts* as survival, relapse, improvement, worsening, or death. These predictions depend on whether there is a temporal *restriction* related to the prediction or not. Temporal restrictions may be represented as a single point (e.g. probability of suffering a relapse "after one year") or as multiple independent points in time [4] (e.g. probability of getting an improvement "within the next three months"). In [3], *prognostic models* are classified into those that predict on populations (e.g. patients that are in a similar condition) and those others that predict on individuals. An additional feature of the above methodologies is whether they are able to predict only one fact (e.g. survival) or whether they are able to predict several facts simultaneously.

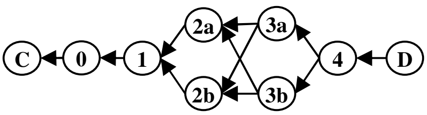
A feasible approach to obtain predictions on several facts simultaneously is based on the concept of *patient condition*, which represents the state of the patient concerning a disease. Thus, finding out the probability of a patient to cure, to

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 489–499, 2007.

improve, to worsen, or to die is equivalent to calculate how likely it is that this patient evolves from his current condition to a condition representing cure, a better than the current condition, an equivalent to the current condition, a worse condition, or the death condition, respectively.

All the possible patient conditions (i.e. states) of a disease define an order relation that represents the pair-wise comparison of the *severity* of the possible conditions in the disease. So, for instance in breast cancer, stage 4 (patients with metastasis) represents a patient condition that is worse than stage 1 (where the tumour is less than 2 cm across and it is not spread). Unfortunately, the severities of two patient conditions are not always comparable or, if they are comparable, it is not always possible to establish one as clearly better than the other one. Therefore, the relationships among the patient conditions of a disease in health-care are frequently represented with *partial orders* which for complex diseases as cancer they are created after an agreement between experts. However, the so created partial orders are not necessarily designed to represent conditions and relationships from a point of view of the severity of the disease but, for instance, to represent the relationships among these conditions from a practical point of view like the sort of recommended treatment is. This can foster differences between what the theoretical model represents (i.e. the expert-based partial order or *standard partial order*) and what is really observed at the health-care centres (i.e. the experience-based partial order). For example, for the data of the SEER repository [7] describing real breast cancer cases, it is observed that 15% of these cases are in a condition whose severity does not correspond to the severity of the stage indicated by the TNM Staging System [8] in Fig. 1.

The reason for that is that the degree of severity of a particular patient condition is not necessarily based on whether this patient fulfils a set of facts or not, but on the combination of the degrees of severity of each one of the variables that define the state of a patient in a particular disease. For instance, it does not seem very wise to admit patients with breast cancers of 2.0 cm in stage 2 (i.e. severity 2), and at the same time do not consider the possibility of a patient with a 2.1 cm tumour to be in stages with severities below or equal to 2 just because the definition of stage 2 in breast cancer sets the size upper limit in 2 cm. Following with the example, it could be the case that the first patient with a 2 cm tumour has other complications affecting the seriousness of his disease, making his condition more severe than the one of the second patient, and causing the prognostic of the first patient not to be very accurate.



TNM BREAST CANCER CONDITIONS

Stage C: no breast cancer observed (Cure). **Stage 1:** The tumour size <2 cm; armpit lymph nodes not affected; cancer not spread. **Stage 2a:** no cells in lymph nodes; cancer in outer covering of the bowel. **Stage 2b:** cancer in outer covering of bowel wall & in nest tissues/organs, lymph nodes not affected; cancer not spread. **Stage 3a:** cancer in inner layer of bowel wall or in the muscle layer; 1-to-3 nearby lymph nodes contain cancer cells. **Stage 3b:** cancer through the bowel wall or in surrounding body tissues/organs; 1-to-3 nearby lymph nodes with cancer cells. **Stage 4:** any size; armpit lymph nodes can be affected; metastasis to other parts of the body. **Stage D:** The patient died (Death).

Fig. 1. TNM Staging System for Breast Cancer

In order to support the correct joint analysis of the condition of a patient with respect to both the standard partial order and the experience-based partial order, it is required to develop algorithms to derive partial orders from the patient records stored in hospital databases. The purpose of this is twofold: on the one hand, these algorithms can be used to generate new health-care knowledge on the feasible stages of a particular disease, and on the other hand, they can be combined with probability theory to increase the accuracy of prognosis on the evolution of a patient.

This paper describes an algorithm to induce partial orders on the patient conditions of a disease. The induction process takes the data of the patients that are registered in the hospital databases and that are described in terms of the variables that condition the health state of the patient in the target disease, and produces a partial order that, together with a state-transition diagram that represent the changes of condition of the patients in the healthcare centre, is able to predict the evolution of new patients.

The rest of the paper has four sections. Section 2 formalises the problem and proposes the structures that the algorithm in section 3 uses to induce partial orders on the feasible patient conditions of a disease. Section 4 describes the tests and the results of these algorithms on three sorts of cancer. The conclusions of the work are exposed in section 5.

2 Condition-Based Prognosis

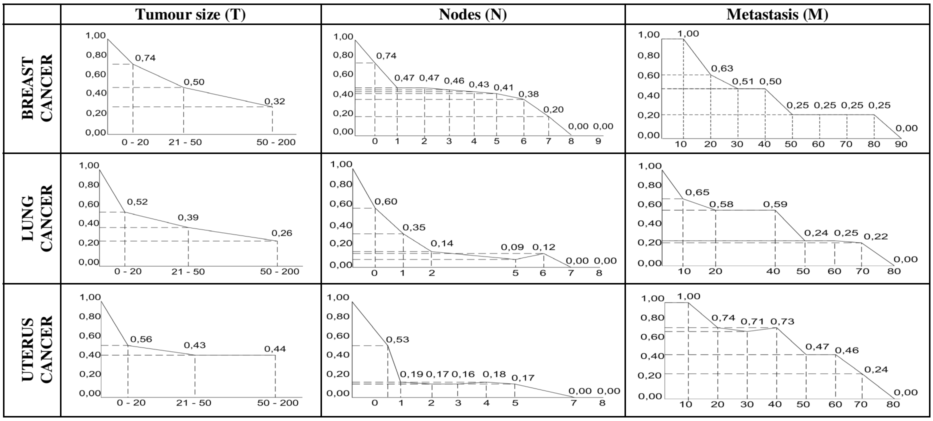
In the process of making a prognosis about the evolution of the health of a patient within a probabilistic framework, there are three main questions to be answered: what are the possible conditions of a patient in the selected disease?, what sort of order there is to compare the seriousness of these conditions?, and how the past evolutions registered in the hospital databases can be used to define a probabilistic model to support the prognostic process?

2.1 Detecting Disease Conditions

For each particular disease $\mathbb{D}$, there is a set of descriptive variables V=$\{v_1, ..., v_k\}$ with respective domains $Dom(v_i)$; i=1, ..., k. Each variable v_i represents a property of the disease that is relevant to understand the condition of the patients suffering from that disease. Each v_i defines a *severity* function s_i: $Dom(v_i)\rightarrow[0,1]$ that provides the degree of seriousness of each one of the values that the variable can take. That is to say, $s_i(v)$ is a value between zero and one representing the severity of the condition of any patient for which v_i takes the value v, zero being the lowest severity (i.e. null), and one being the highest one. *Slightness* is defined as the opposite of severity, i.e. $\mu_i(v)$=1- $s_i(v)$. For the sake of optimism, the rest of the paper will be based on the concept of slightness rather than on severities. So, Table 1 contains the slightness functions for the variables of tumour size (T), nodes (N) and metastasis (M) in the breast, lung and uterus cancer. These functions are derived from the information contained in the SEER database [7] and may vary from other sources of information.

Given a set of variables V, the condition of a patient p (or *patient condition* c_p) can be formally described as an element of the set $Dom(v_1)\times Dom(v_2)\times\ldots\times Dom(v_k)$ (i.e. c_p=$(a_1, \ldots, a_k)$, a_i being the value p has for variable v_i), and the *global slightness* of c_p

Table 1. Slightness functions for the variables *T*, *N* and *M* in the domains of Breast Cancer, Lung Cancer, and Uterus Cancer



in the disease D as a combination of all the slightness functions of the descriptive variables. Many sorts of combinations exist [1], though here only the arithmetic mean is used. So, $\mu(c_p)=1/k \cdot \Sigma_i\ \mu_i(a_i)$ is the function to calculate the global slightness of any patient condition with values a_1, …, a_k in the variables of V. This combination is possible since a correlation analysis of the data in the SEER database shows that T, N and M are mutually independent variables. Although they are not considered here, alternative combination functions should be taken if the variables to combine are not independent.

A patient condition of a disease $\mathbb{D}$ (or *disease condition* C) is defined as a restriction on the domains of the variables of that disease. So, any disease condition can be formalised as C=(D_1, ..., D_k) with $D_i \subseteq Dom(v_i)$, i=1, ..., *k*, and represents a common state of a set of patients suffering from $\mathbb{D}$. The set of all the disease conditions C_1, …, and C_n of a disease $\mathbb{D}$ contains the alternative states in which a patient of that disease can be.

For some diseases the set of disease conditions C_i are fixed and well defined, like in cancers where the *Tumour Node Metastasis Staging System* (TNM) [8] was created by the American Joint Committee on Cancer (AJCC) to describe the alternative conditions of diverse cancers; for example, the stages 0, 1, 2a, 2b, 3a, 3b, and 4 in breast cancer that Fig. 1 extends with the extreme conditions *cure* (left side C node) and *death* (right side D node).

In other diseases where there in not an agreed criterion on the set of conditions, these can be obtained from the application of a non-supervised clustering algorithm on a representative sample of *patient conditions* described in terms of the set of variables V. Two alternative sorts of clustering algorithms can be applied: data clustering and conceptual clustering. Data clustering algorithms like *kMEANS* [5] obtain clusters of similar patient conditions that are dissimilar to the patient conditions in other clusters. On the contrary, conceptual clustering algorithms like *COBWEB* [2] obtain clusters as expressions describing the patient conditions contained in the cluster, in terms of the variables in V.

The application of a clustering algorithm can be made directly on the values of the variables in V (i.e. patient respective values a_1, …, a_k) or, alternatively, on the values of the slightness functions of the variables in V (i.e. values $\mu_1(a_1)$, …, $\mu_k(a_k)$). Whereas the first option puts patient conditions with similar descriptions in the same cluster, the second group of algorithms gathers patient conditions with similar slightness values in the same cluster.

2.2 Using Partial Orders to Sort the Seriousness of the Disease

The global slightness function μ defines a complete order relation among the patient conditions that can be described in terms the variables in V. So, for any particular disease, if c_i and c_j represent two patient conditions and $\mu(c_i)>\mu(c_j)$, we interpret that c_i is better than c_j. Nevertheless, this sort of order relation cannot be extended to the comparison of disease conditions where two conditions C_i and C_j of the same disease can not only represent one a worse state than the other, but also incomparable states from the point of view of their respective slightness. This implies that, for any disease D, the order relation of the feasible disease conditions is not necessarily complete.

Formally, given a set of elements A, a *partial order* $P \subseteq A \times A$ on these elements is a binary relation such that P is reflexive (i.e. $e_i \in A \Rightarrow (e_i, e_i) \in P$), anti-symmetric (i.e. $(e_i, e_j) \in P$ and $(e_j, e_i) \in P \Rightarrow e_i = e_j$), and transitive ($(e_i, e_j) \in P$ and $(e_j, e_k) \in P \Rightarrow (e_i, e_k) \in P$).

Partial orders are typically represented as directed acyclic graphs where all the edges that are deducible by transitivity (i.e. *weak* relations) are omitted.

A set of disease conditions $\{C_1, ..., C_n\}$ on a disease D defines a partial order. This partial order can be used to know whether one condition is better or worse than other condition, or if they cannot be compared. For example, Fig. 1 depicts a directed acyclic graph that represents the standard partial order of the breast cancer conditions according to the TNM staging system [8]. It shows, for instance, that a patient in stage 2a is healthier than one patient in stage 3a or 3b (direct edge connection), or 4 (connected by edge transitivity), and not comparable in terms of slightness to patients in stage 2b.

The difference between two partial orders P_1 and P_2 can be measured in terms of the cardinality of the set $(P_1 \cup P_2) - (P_1 \cap P_2)$.

2.3 Using State-Transition Diagrams to Represent the Cases in Hospital DBs

In the previous section we showed how the conditions of a disease define a partial order of their respective slightness. This conceptual structure, however, is unable to represent the evolutions of patients in time which are based on patient improvements, worsenings and stable periods. *State-Transition Diagrams* are directed graphs that model behaviours in terms of states, transitions and actions. Here, states stand for the conditions of a disease, transitions are the evolutions of the observed patients as their conditions change in time, and actions remain unused. Formally speaking, if $\mathbb{C}$ is a set of disease conditions of a disease $\mathbb{D}$, a state-transition diagram is a pair ($\mathbb{C}$, t) such that t: $\mathbb{C} \times \mathbb{C} \rightarrow \mathbb{N}$ is the transition function that, for each couple of disease conditions

C_i and C_j in $\mathbb{C}$, $t(C_i,C_j)$ is the number of patients whose conditions evolve directly from C_i to C_j. The *inflow* and the *outflow* of a disease condition C_i can be calculated with the functions $in(C_i)=\Sigma_j\ t(C_j,C_i)$ and $out(C_i)=\Sigma_j\ t(C_i,C_j)$, respectively.

If this model is used to represent the evolutions of a set of patients across the feasible conditions of a disease, it must be extended with the *admission* and the *discharge functions* a: $\mathbb{C} \rightarrow \mathbb{N}$ and d: $\mathbb{C} \rightarrow \mathbb{N}$ such that for any condition C_i, $a(C_i)$ is the number of patients arriving in condition C_i, and $d(C_i)$ the number of patients leaving from (or still remaining in) condition C_i. See that, for any disease condition C_i, $a(C_i)+in(C_i)$ must be equal to $out(C_i)+d(C_i)$. Then, if $n_i=a(C_i)+in(C_i)$ represents the number of times any patient has been in condition C_i, and $n_t=\Sigma_i\Sigma_j\ t(C_i,C_j)$ the number of changes of disease condition of all the patients registered in a hospital database, the probability of a patient to be in condition C_i is $p(C_i)= n_i/n_t$, the probability of a patient p in condition C_i to evolve to C_j in one transition is $p(C_i,C_j)= t(C_i,C_j)\ /\ n_i$, and the probability of finding a patient that evolves from C_i to C_j is $t(C_i,C_j)\ /\ n_t$.

The above function $p(C_i,C_j)$ can be used to compute the probability of a patient to evolve from one set of disease conditions $\mathcal{A} \subseteq \{C_1, \ldots, C_n\}$ to another set of disease conditions $\mathcal{B} \subseteq \{C_1, \ldots, C_n\}$ in one step as $Pr(\mathcal{A}, \mathcal{B}) = \Sigma_{Ci\in\mathcal{A}}\ \Sigma_{Cj\in\mathcal{B}}\ p(C_i,C_j)$. In its turn, this function, together with a partial order P on the disease conditions, can be used to make prognoses on the likelihood a patient gets cured, improves, worsens, dies, or survives. See equations 1 to 5, respectively where *Condition(p)* represents the current condition of the patient, *cure* is the condition of a healthy patient, and *death* is the condition representing a deceased patient.

$$Pr(p \text{ cures}) = Pr(\{Condition(p)\},\{\text{cure}\}) \tag{1}$$

$$Pr(p \text{ improves}) = Pr(\{Condition(p)\},\{C\text{: } (Condition(p),C)\in P\}) \tag{2}$$

$$Pr(p \text{ worsens}) = Pr(\{Condition(p)\},\{C\text{: } (C, Condition(p))\in P\}) \tag{3}$$

$$Pr(p \text{ dies}) = Pr(\{Condition(p)\},\{\text{death}\}) \tag{4}$$

$$Pr(p \text{ survives}) = 1\text{-}Pr(p \text{ dies}) \tag{5}$$

3 Induction of Partial Orders

Condition-Based Prognosis as it was introduced in section 2 is a three step process that starts with the determination of the conditions of a disease (here, we will consider the set of conditions already available). Once the disease conditions are fixed, a second step takes the data of the evolutions of patients in a health-care centre to induce both a partial order on these conditions, and also a state-transition diagram that contains the probabilities $p(C_i,C_j)$ of evolving from any disease condition C_i to any other disease condition C_j in the context of the selected health-care centre. After that, a third step can be applied that consists on the utilisation of both structures to predict the evolution of new patients: the partial order provides the semantic meaning of what “cure”, “improve”, “worsen”, “die”, or “survive” means in the context of the patient current medical condition, and the state-transition diagram supplies the probabilities needed to compute the final prognostic value. This section describes the procedures to carry out the second and the third steps.

3.1 The Data Model

The two main structures used in condition-based prognosis (i.e. partial order and state-transition diagram) are generated from the same database. This database contains the data about the evolutions of the patient conditions in a health-care centre. The basic data structure is the episode of care. An *episode of care* (EOC) contains all the medical information about the treatment of one patient between the date of admission and the date of discharge. In our approach, an EOC is represented as a sequence of patient-professional *encounters* in which the professional observes the condition of the patient and proposes a course of action. Formally, if V=$\{v_1, ..., v_k\}$ is a set of descriptive variables of the patient conditions in a disease $\mathbb{D}$ and $\mathbb{A}=\{a_1, .., a_p\}$ is a set of medical actions, then an encounter e is a pair (c, a) such that c is a patient condition (i.e. $c \in Dom(v_1) \times Dom(v_2) \times \ldots \times Dom(v_k)$) and a is a subset of actions in A; an EOC is a sequence $e_1, ..., e_q$ of encounters, and the database is a list of EOCs.

3.2 The Statistical Model

According to the data structure described above, for any pair of disease conditions (C_i, C_j), we can apply a statistical procedure to determine, in a first stage, whether there is an order relation between C_i and C_j and, if there is one, in a second stage, decide which of the two conditions represents a better state of the disease from a health point of view (i.e. the order of the relation between C_i and C_j). Once all the pairs of disease conditions are considered, a *statistically significant partial order* on these conditions is obtained. Here, the above mentioned two stages are implemented as statistical hypothesis Student's t-tests.

In the study of a disease D, with $\{C_1, \ldots, C_n\}$ the set of all possible conditions of D, and provided a database containing a representative sample of encounters of all the patients that have been treated of that D, the description of the state of the patient in each encounter e_k in terms of the variables in V defines a patient condition c_k with a slightness value $\mu(c_k)$ –or $\hat{\mu}(c_k)$ in statistics notation. Simultaneously, this patient condition c_k classifies the encounter in one of the disease conditions $C_1, \ldots, C_n$.

Let us call E_k the set of the encounters in the database that are classified in C_k, and $S_k=\{\mu(c_j)\text{: } e_j \in E_k\}$ the set of μ-values of their patient conditions. Then, for any pair of disease conditions C_i and C_j, the respective sets S_i and S_j are the two independent samples of a Student's t-test with null hypothesis the means of the slightness values of the elements in C_i and the elements in C_j are equal, provided that the underlying distributions are normal.

Only if the null hypothesis is rejected, C_i and C_j have an order relation whose sense is evaluated with a new Student's t-test with null hypothesis the means of the slightness values of the elements are larger in C_i than in C_j. Both t-tests are based on the t-value (6) where μ's, σ's and n's represent the mean, standard deviation, and number of elements of the samples, respectively.

$$\beta = \frac{\overline{\mu_i} - \overline{\mu_j}}{\sqrt{\frac{\sigma_i^2}{n_i - 1} + \frac{\sigma_j^2}{n_j - 1}}} \tag{6}$$

3.3 The Algorithm

An algorithm to induce partial orders under the previously described statistical model is introduced in this section. This algorithm realizes the induction process according to the data and the statistical models of sections 3.1 and 3.2, respectively. The final result of the algorithm is a partial order that explains the slightness degree of a disease in terms of the improvement or worsening between the conditions of a disease.

```
Algorithm MakePartialOrder (C, data, $\alpha$)
{Let C = {C1,…,Cn} be a set of conditions on a disease D}
{Let data    = {$EOC_1$, …, $EOC_k$} be a list of episodes of care of D}
{     $EOC_i$   = {$e_{i1}$, …, $e_{iki}$} the list of encounters in $EOC_i$, i=1..k}
{Let $\alpha$ the statistical significance of the test -e.g. 0.01}
   j: float
   PO = $\varnothing$; {empty partial order on the set of disease conditions C}
   For any pair of conditions (Ci, Cj) in C×C
       Ei = {$e_{xy} \in \cup_z EOC_z$: Ci is the condition of the patient in encounter $e_{xy}$}
       Ej = {$e_{xy} \in \cup_z EOC_z$: Cj is the condition of the patient in encounter $e_{xy}$}
       Si = {$\mu(c_x)$: $c_x$ is the condition of the patient in $e_x$, for all $e_x \in$Ei}
       Sj = {$\mu(c_x)$: $c_x$ is the condition of the patient in $e_x$, for all $e_x \in$Ej}
       Calculate the t-value $\beta$ according to equation 6
       If $|\beta| < t_{\alpha/2}$ (first hypothesis test indicates Ci and Cj are related) then
         If $\beta > t_\alpha$ (second hypothesis test indicates Ci is better than Cj) then
           Insert (Ci,Cj) in PO;
         else
           Insert (Cj,Ci) in PO;
       End If; End If;
   End For;
   Write the order relation PO;
End Algorithm.
```

4 Experiments

In order to induce partial orders, we used the databases on the diseases Breast Cancer (55939 encounters), Lung Cancer (19491 encounters) and Uterus Cancer (705 encounters) obtained from the SEER Cancer Incidence Public-Use Database [7]. These databases contain information on patient conditions based on three variables: Tumour Size, Lymph Nodes, and Metastasis classified according to the TNM System [8]. Data with unknown or missing values are removed from the databases. The distribution of these data according to each disease condition is described in Table 2.

Table 2. Distribution of episodes according to each disease condition

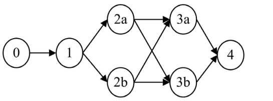
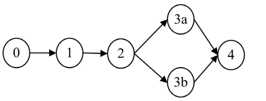
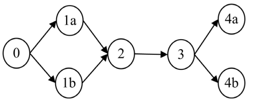
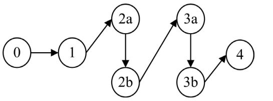
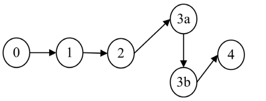
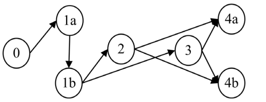
Cancer Disease	Disease conditions									Total
	0	1a	1b	2a	2b	3a	3b	4a	4b	
Breast	7073	25566		13387	6550	1456	940	967		55939
Lung	11	7298		1338		2629	3022	5193		19491
Uterus	51	242	203	79		45		5	80	705

Two sorts of tests have been performed on these databases: one that is used to compare the difference between the standard partial orders which are proposed by the TNM Staging System [8], and the experience-based partial orders obtained by the inductive algorithm introduced in section 3.3 when it is applied on the proposed databases. The second test is about how these differences affect the process of prediction on the facts of cure, improvement, worsening, death, and survival in breast, lung, and uterus cancers.

4.1 Results on the Induction of Partial Orders

Table 3 shows both the standard partial orders [8] and the partial orders the proposed algorithm induces form the three databases. The distances between the standard and the induced partial orders are 2, 1 and 2, respectively. These differences are caused either by the detection of new relations that were not present in the standard partial order or by the elimination of relations that do not achieve the statistical significance level required to be part of the experience-based partial order. So in breast cancer, the relations 2a-2b and 3a-3b are statistically justified though they were not in the standard partial order. A similar case is observed in lung cancer with the relation 3a-3b, and in uterus cancer with relation 1a-1b. In this last domain, the SEER database does not provide enough evidence to keep the standard order relation between stages 2 and 3 in the experience-based partial order.

Table 3. Partial orders induced

	Breast Cancer	Lung Cancer	Uterus Cancer
Standard Partial Order			
Experience-Based Partial Order			

These single differences between standard and experience-based partial orders are cause of new differences when the transitivity property is applied, and the final differences increase to 3%, 2%, and 10% of the total number of binary relations, this meaning that 3, 2, and 10 out of 100 comparisons get different responses whether the standard or the experience-based partial orders are queried.

4.2 Results on the Condition-Based Prognosis

Equations 1 to 5 in section 2.3 are used to calculate the probabilities of improvement, worsening, cure, death and survival in Breast, Lung and Uterus cancers for both, the standard partial order, and the experience-based partial order the algorithm in section 3.3 obtains for the data of the SEER repository [7], representing real patients.

	BREAST CANCER							STND		EXP.B	
	0	**1**	**2a**	**2b**	**3a**	**3b**	**4**	**I**	**W**	**I**	**W**
0	1	0	0	0	0	0	0	0	0	0	0
1	0.6	0.2	0.1	0.1	0	0	0	0.75	0.25	0.75	0.25
2a	0.3	0.2	0.1	0.2	0.1	0.1	0	0.64	0.36	0.55	0.44
2b	0.1	0.1	0.2	0.1	0.1	0.2	0.2	0.36	0.64	0.44	0.55
3a	0	0	0.1	0.1	0.3	0.2	0.3	0.39	0.61	0.29	0.71
3b	0	0	0	0.2	0.3	0.1	0.4	0.25	0.65	0.55	0.44
4	0	0	0	0	0	0	1	0	0	0	0

	LUNG CANCER						STND		EXP.B	
	0	**1**	**2**	**3a**	**3b**	**4**	**I**	**W**	**I**	**W**
0	1	0	0	0	0	0	0	0	0	0
1	0.8	0.1	0.1	0	0	0	0.88	0.11	0.88	0.11
2	0.1	0.3	0.1	0.2	0.2	0.1	0.44	0.55	0.44	0.55
3a	0	0.1	0.2	0.1	0.2	0.4	0.39	0.61	0.33	0.66
3b	0	0	0.1	0.3	0.1	0.5	0.24	0.76	0.44	0.55
4	0	0	0	0	0	1	0	0	0	0

	UTERUS CANCER							STND		EXP.B	
	0	**1a**	**1b**	**2**	**3**	**4a**	**4b**	**I**	**W**	**I**	**W**
0	1	0	0	0	0	0	0	0	0	0	0
1a	0.4	0.3	0.1	0.1	0.1	0	0	0.68	0.32	0.57	0.43
1b	0.4	0.4	0.1	0	0.1	0	0	0.75	0.25	0.88	0.11
2	0.1	0.1	0.1	0.1	0.2	0.2	0	0.33	0.67	0.58	0.42
3	0	0	0.1	0.1	0.1	0.2	0.5	0.22	0.78	0.32	0.68
4a	0	0	0	0.1	0.1	0.2	0.6	0.25	0	0.25	0
4b	0	0	0	0	0	0	1	0	0	0	0

Fig. 2. Probabilities of evolution among disease conditions in breast, lung, and uterus cancers

In order to analyse the differences between the prediction values obtained with the utilisation of either the standard or the experience-based partial orders, the probabilities $p(C_i,C_j)$ that are obtained from the real evolution of a set of patients, are used to define a matrix of patient evolutions. Fig. 2 shows the probability matrices employed to analyse these differences in the cases of breast, lung, and uterus cancers.

The probabilities of cure, death, and survival are identical for the standard and the experience-based partial orders, as expected, since the conditions of *cure* and *death* are the same in both partial orders. However, the predictions on improvement (I) and worsening (W) differ if we use one or the other partial orders, as the numbers in grey indicates. Some of these differences cause the prognostic with the standard partial order to provide excessive "hope" (e.g. in uterus cancer, patients in stage 1a are given 68% of improvement, whereas the experience says that only 57% will improve), or excessive "despair" (e.g. in uterus cancer, patients in stage 2 get 67% of worsening, when reality shows that it is only 42%).

5 Conclusions

In this paper, we have introduced a method to induce partial orders for patient conditions in a disease, which is part of a broader work in the area of machine learning to support healthcare activities [6]. Here, the partial orders which are built from real experiences happened in health-care centres show the gap there is between the criteria to assess the patient condition proposed by medical experts (standard partial order), and the criteria coming out of the medical daily situations (experience-based partial order).

From the tests described in the previous section, we can conclude there are clear structural differences between the standard partial orders proposed by the physicians and those others that are induced from the data of the SEER repository about real patients. A direct implication of these differences is that the prognosis about the evolution of patients may change drastically. This effect has been confirmed with the results of the tests performed which may drive the physician to incorrect predictions of patient future improvements and worsenings.

Acknowledgment. This work has been partially funded by the European IST Project K4CARE (IST-026968) and the Spanish Project HYGIA (TIN2006-15453-c04).

References

1. Figueira, J., Greco, S., Ehrgott, M. (ed.): Multiple Criteria Decision Analysis. State of the Art Surveys. Springer's International Series, New York (2005)
2. Fisher, D.: Knowledge acquisition via incremental conceptual clustering. Machine Learning 2, 139–172 (1987)
3. Lucas, P., Ameen, A.-H. (ed.): Prognostic Methods in Medicine. Artificial Intelligence in Medicine vol. 15, pp. 105–119 (1999)
4. Machado, O.L.: Methodological Review: Modelling Medical Prognosis: Survival Analysis Techniques. Journal of Biomedical Informatics 34, 428–439 (2001)

5. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Procs of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1st edn., pp. 281–297. University of California Press, Berkeley (1967)
6. Riaño, D., Bohada, J.A., Welzer, T.: The DTP model: Integration of intelligent techniques for the decision support in Healthcare Assistance. EIS2004 (2004)
7. SEER Cancer Statistics Review. Surveillance, Epidemiology, and End Results (SEER) program public-use data (1973-2003). National Cancer Institute, Surveillance Research program, Cancer Statistics Branch, released April 2006, based on the (November 2005 submission) http://www.seer.cancer.gov
8. Sobin, L.H., Wittekind, C.: TNM Classification of Malignant Tumours, 6th edn. John Wiley & Sons, New Jersey (2002)