ELSEVIER

Artificial Intelligence in Medicine 26 (2002) 87–107

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Logistic-based patient grouping for multi-disciplinary treatment

Laura Măruşter[a,*], Ton Weijters[a], Geerhard de Vries[a,d], Antal van den Bosch[b], Walter Daelemans[b,c]

[a]*Department of I&T, Faculty of Technology Management, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*
[b]*ILK/Computational Linguistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands*
[c]*CNTS Linguistics, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium*
[d]*Prismant Institute for Health Care Management, P.O. Box 9697, 3506 GR Utrecht, The Netherlands*

Received 5 March 2002; accepted 12 March 2002

Abstract

Present-day healthcare witnesses a growing demand for coordination of patient care. Coordination is needed especially in those cases in which hospitals have structured healthcare into specialty-oriented units, while a substantial portion of patient care is not limited to single units. From a logistic point of view, this multi-disciplinary patient care creates a tension between controlling the hospital's units, and the need for a control of the patient flow between units. A possible solution is the creation of new units in which different specialties work together for specific groups of patients. A first step in this solution is to identify the salient patient groups in need of multi-disciplinary care. Grouping techniques seem to offer a solution. However, most grouping approaches in medicine are driven by a search for pathophysiological homogeneity. In this paper, we present an alternative logistic-driven grouping approach.

The starting point of our approach is a database with medical cases for 3603 patients with peripheral arterial vascular (PAV) diseases. For these medical cases, six basic logistic variables (such as the number of visits to different specialist) are selected. Using these logistic variables, clustering techniques are used to group the medical cases in logistically homogeneous groups. In our approach, the quality of the resulting grouping is not measured by statistical significance, but by (i) the usefulness of the grouping for the creation of new multi-disciplinary units; (ii) how well patients can be selected for treatment in the new units. Given a priori knowledge of a patient (e.g. age, diagnosis), machine learning techniques are employed to induce rules that can be used for the selection of the patients eligible for treatment in the new units. In the paper, we describe the results of the above-proposed methodology for patients with PAV diseases. Two groupings and the accompanied classification rule sets are presented. One grouping is based on all the logistic variables, and another

* Corresponding author. Tel.: +31-40-2473703; fax: +31-40-2432612.
E-mail address: l.maruster@tm.tue.nl (L. Măruşter).

0933-3657/02/$ – see front matter
PII: S0933-3657(02)00054-4

grouping is based on two latent factors found by applying factor analysis. On the basis of the experimental results, we can conclude that it is possible to search for medical logistic homogenous groups (i) that can be characterized by rules based on the aggregated logistic variables; (ii) for which we can formulate rules to predict to which cluster new patients belong.

Keywords: Patient classification systems; Clustering; Rule induction; Logistic management

1. Introduction

In The Netherlands, as in many other countries in the world, there is a markedly growing demand for the coordination of patient care. Strong emphasis is placed on medical and organizational efficiency and effectiveness to control national healthcare expenditures. One of the recognized efficiency problem is that sub-optimally coordinated care often results in redundant and overlapping diagnostic procedures performed by medical specialists from different specialties within the same hospital. Coordination becomes especially important when hospitals structure their healthcare into specialty-oriented units, and care for patients is not constrained within single units. From a logistic point of view, this creates a tension between the control over the units, and the coordination needed among units to control the patient flow.

The total flow of the patients in a hospital can be divided into mono- and multi-disciplinary patients. Multi-disciplinary patients require the involvement of different specialties for their medical treatment. Naturally, these patients require more efforts regarding the coordination of care. A possible solution is the creation of new multi-disciplinary units, in which different specialties coordinate the treatment of specific groups of patients. A first step in this solution is to identify salient patients groups in need of multi-disciplinary care. Furthermore, adequate selection criteria must exist to select new patients for treatment in a multi-disciplinary unit. As we will demonstrate, grouping and classification techniques seem to offer a solution.

In the medical domain, various grouping and classification techniques are developed and used [3,6]. They can be categorized by their purposes as utilization, reimbursement, quality assurance and management applications [12]. For example, Fetter's Diagnostic Related Groups (DRGs) [6] and their refinements [7] are homogeneous in terms of use of resources, but the elements within a single group show rather high variability and low homogeneity from the underlying process point of view [16]. Starting from the original DRG concept, researchers and professionals organized themselves into a joint network for providing efficient methods for health management at different levels of care under the name of *case-mix classification systems* [3]. However, none of the existing classification systems are homogeneous from the underlying logistic process point of view [16]. A solution will be to consider a logistic classification system that results in a higher logistic homogeneity of groups.

In this paper, we investigate the possibility of building an alternative, logistic-driven grouping and classification system for medical multi-disciplinary patients with the aid of machine learning techniques.

In the medical domain, machine learning methods are used successfully for diagnostic, prognostic, screening monitoring, therapy support purposes [8,9], but also for overall patient management tasks like planning and scheduling [11,13].

In the present work, we combine unsupervised and supervised machine learning techniques to achieve our three-fold objectives:

(i) First, we want to be able to classify patients in groups that are homogeneous from the underlying process point of view. For this purpose, we operationalize the concept of logistic complexity into different aggregate logistic variables that will be used further in clustering. Subsequently, we characterize the obtained clusters by rules based on the aggregated logistic variables.
(ii) Second, we aim at developing a rule predictive model that can assign a new patient on the basis of some given personal information (age, gender, chronic diagnosis), to the most suitable logistic group instantly. Thus, the a posteriori information encapsulated in the aggregated logistic variables will be used for the development of homogeneous logistic clusters; conversely, a priori personal information will be used to assign new patient as soon as possible to a cluster.
(iii) Third, we illustrate how machine learning (data mining) techniques can aid in this process.

We plan to assess the quality of the logistic clusters considering a combination of different criteria. First, we want our obtained clusters to be logistically *homogeneous*. Second, both the cluster characterization rules and the predictive rules should make sense from the medical point of view; thus we are interested on the *intelligibility* and *usefulness* of our rules.

The structure of the rest of the paper is in line with the general knowledge discovery framework as proposed by Cios et al. [4]. In Section 2, we describe the problem domain. We provide a medically-oriented description of the multi-disciplinary patients investigated in this study, all treated for peripheral arterial vascular (PAV) diseases. From the logistic point of view, we then elaborate on the importance of the underlying processes of medical multi-disciplinary patients, particularly when one aims to optimize the patient throughput. In Section 3, we describe the collection and preparation of data and the operationalization of the logistic complexity concept. Section 4 describes the clustering experiments for finding logistically homogeneous groups. Our approach of developing predictive models is presented in Section 5. In Section 6, we discuss the results of the used data mining techniques. Finally, in Section 7 we formulate the conclusions on the basis of our current findings, and describe some future research.

2. Understanding the problem domain

2.1. The medical problem domain

Patients who require the involvement of different specialties are hardly a new phenomenon in healthcare. In general, one can say that because of the increasing specialization of doctors within the hospital and an aging population this group of patients is increasing.

Recent studies in The Netherlands show that approximately 65% of the patients visiting a hospital are multi-disciplinary [17]. Consequently, certain special arrangements have emerged for these patients. For instance, some hospitals have special centers in which different specialties work together on backbone problems.

Patients with PAV diseases (peripheral refers to the entire vascular system except for the heart and brain) are a good example of multi-disciplinary patients. Surgery, internal medicine, dermatology, neurology and cardiology are the specialties most frequented involved by the treatment of these patients. Alarmingly, a recent study of The Netherlands Heart Foundation shows that the care for these patients leaves much to be desired, because it is too dispersed: it is difficult for doctors in primary healthcare to know what specialty to refer to; knowledge within the hospital is dispersed; there is a lack of within-hospital cooperation; and there are impediments to scientific research.

Arguably, one important reason for these problems is that patients with PAV diseases are grouped on the basis of medical homogeneity, in the hope that this will result in logistically homogenous groups. However, PAV are a variety of diseases, both acute and chronic, life-threatening, or invalidating. Table 1 illustrates that describing these diseases as a group is complex. One complaint can have many different causes, one cause can have different manifestations and there is complexity in cause and effect between the pathologies.

One of the consequences of the complexity of expressing these patients in medical terms is that the homogeneity of the underlying treatment processes of these patients is low. This leads us to the logistic perspective of our approach.

2.2. The logistic problem domain

In this subsection, we expound our view on the logistics and subsequently we set up our logistic goals. Logistics is defined as "the coordination of supply, production and distribution process in manufacturing systems to achieve a specific delivery flexibility and delivery reliability at minimum costs" [1,15]. Translated to healthcare organizations, it comprises the design, planning, implementation and control of coordination mechanisms between patient flows and diagnostic and therapeutic activities in health service organizations. The goal is to maximize output/throughput with available resources, taking into account different requirements for delivery flexibility (e.g. differentiating between

Table 1
Patients with PAV diseases expressed in medical terms

Pathologies	Intermediate stage	Manifestation	Measurable and visible symptoms/ complaints	Irreversible disorders and diseases
Arteriosclerosis	Plaque thrombus	Ischaemia	Pain in legs	Impair of organs, muscles and arteries
Disturbed composition of the blood	Plaque thrombus	Ischaemia	Pain in chest	Impair of organs, muscles and arteries
Disturbed metabolism	High concentration of glucose in blood	Insufficient supply of glucose in cells	Fatigued, perspiration, tremble	Disorder of arteries affection of nerves

elective/appointment, semi-urgent, and urgent delivery) and acceptable standards for delivery reliability (e.g. determining limits on waiting list length and waiting times) and acceptable medical outcomes [14,17].

First of all, a production control approach to hospitals requires knowledge about processes. However, the main characteristic of hospital products is that they are organized by specialty: internal medicine, cardiology, pulmonology, etc. The physicians belonging to a specialty are specialized in treating complaints in a well-defined part of the human body; often there are even sub-specializations within a specialty, for instance diabetics, enterology and oncology as specializations within internal medicine. However, from a logistic point of view we are looking for homogeneity of the underlying processes. With this we mean the sequence, timing and execution of activities for patients by the hospital staff (specialists, nurses and paramedics). Distinguishing logistically homogeneous groups appears to be important, because every logistic group can require its own optimal control system. Subsequently, in the following sections we will investigate whether such logistic groups can be found in reality.

3. Data collection and preparation

The two logistic characteristics to typify a production situation, or in this case the care process of a patient are (i) the complexity of the care process; (ii) the routing variability of the care process. In this paper, we concentrate on the first type, i.e. the complexity of the care process of a patient. Keep in mind, we are referring to logistic complexity, which can be something completely different from medical complexity. Before we come to the part of operationalizing the characteristics, a remark on the subject of data gathering in hospitals is in place. The degree of detail in the majority of hospital registrations is high, but the information is normally hidden in different databases. Relevant patient information can be found in clinical databases, outpatient databases, laboratory databases, etc. When all the patient information is gathered, we have the hospital history of the patient.

When planning to do quantitative investigations based on real-data, one has to be aware that ''some critical steps should be followed'' [5]. What we plan to do is cluster analysis in order to find the logistically homogeneous groups. Therefore, we concentrate on the following aspects: (i) data selection; (ii) the attributes (or variables) that should be recorded (measured); (iii) how to deal with missing data.

3.1. Data selection

The first step for data selection is to establish the criteria on which the data will be chosen. Our target is to investigate PAV patients, because they are a good example of complex multi-disciplinary patients. Therefore, we need to specify what we mean with a PAV patient. For this purpose, interviews were held with specialists from the source hospitals, which revealed that certain types of diagnoses point to our PAV patients. These diagnosis resulted in three lists: (a) a list with degenerative underlying chronic diseases; (b) a list with PAV diseases; (c) a list with diagnosis related with chronic or PAV diseases. We selected the whole population of patients who have at least one diagnosis from list (a) or (b)

from the Elisabeth Hospital located in Tilburg, The Netherlands. Note, that we work further with the complete population of PAV patients and not with a sample. The degenerative underlying chronic diseases from list (a) (e.g. diabetes) are the cause of a lot of PAV diseases, therefore together with diagnoses from list (b) they have been considered as selection criteria. For patients with diagnoses from list (a) or (b), all records related to visits in different departments of the hospital were extracted. These records contain information mainly related to:

- personal characteristics: age, gender, address, date of birth and date of death if the patient is deceased, etc.;
- characteristics of the policlinic visit: specialist, date, referral date, referring specialist (if the general practitioner requests the visit or another specialist from the hospital), urgency, etc.;
- characteristics of the clinical admission: specialist, date, diagnosis (one main diagnosis and up to eight possible secondary diagnoses), treatment, referring specialist (if the general practitioner requests the admission or another specialist from the hospital), urgency or planned admission, etc.;
- radiology, functional investigations information, other investigations.

These information fields were used to build a time-ordered history for 3603 patients. Please note that our purpose is not to analyze the underlying processes in the patient's history. For instance, given a patient who breaks a leg in February, and undergoes an appendectomy in August, we find both events in the patient's history, but we do not want to consider the two facts as one medical case. To this end, we established, with the aid of medical specialists, a set of heuristic rules for splitting the patient's history into separate medical cases. We considered only those medical cases that contain at least one clinical admission (because only in case of clinical admission we have recorded the diagnosis). The end result was a database with 4395 records as medical cases of the 3603 considered patients.

3.2. Choice of the variables

As stated before, the goal of our clustering is to find clusters of cases that are homogeneous related to the complexity of the care process. However, the literature does not offer a unique measurement of care process complexity. Based on existing logistic literature concerning complexity, we therefore operationalized the concept of complexity of the underlying process by distinguishing six aggregated logistic variables, each to be investigated as a potential (partial) measurement of care process complexity. We build the six aggregated logistic variables as described below. To illustrate the construction of the logistic variables, we used the following abbreviations: I, internal medicine; C, cardiology; D, dermatology.

1. C_dif_visit: the total number of involved specialties within the medical case. The assumption is that the more specialties are involved, the more complex the medical case is. Suppose that a medical case contains a sequence of visited specialties as follows: I-I-C-I-D-I. Thus, the logistic variable C_dif_visit = 3.

2. C_shift: number of shifts within the medical case, counted by the total number of visits to specialties within the medical case. The assumption is that the more a patient has to go from one specialty to another, counted by the total number of visits, the more complex the medical case.

   As an illustration, let us have a look to the following example. Consider that patient A has a medical case that involves the following sequence of visited specialties: I-I-C-I-D-I; C_shift will be computed as the number of shifts divided with the total number of visits, within the medical case, i.e. $\text{C\_shift}_A = 4/6 = 0.6$. Consider now that patient B has a medical case where the specialties are in the sequence I-I-C-I-I-I-I-D-I-I-I-I-I. Thus, $\text{C\_shift}_B = 4/13 = 0.3$. Obviously, patient A is more complex than patient B, although both A and B ''changed'' specialties four times. Thus, the more a patient has to go from one specialty to another, counted by the total number of visits within the medical case, the more complex the medical case.
3. N_visit_mc: number of visits within the medical case per time-scale. The assumption is that the more visits per time-scale, the more complex the medical case. For example, consider that patient A visited three specialties in 4 weeks, whereas patient B visited three specialties in 12 weeks. Subsequently, $\text{N\_visit\_mc}_A = 3/4 = 0.7$ and $\text{N\_visit\_mc}_B = 3/12 = 0.2$, consequently patient A is more complex than patient B.
4. N_shift_mc: number of shifts within the medical case per time-scale, counted by the total number of visits to specialties. The assumption is that the more shifts per time-scale, the more complex the medical case. For example, consider that patient A has a medical case that involves the following sequence of visited specialties in 14 weeks: I-I-C-I-D-I. Patient B visited the following specialties in 12 weeks: I-I-C-I-I-I-I-D-I-I-I-I-I. Hence, $\text{N\_shift\_mc}_A = 0.6/4 = 0.15$, $\text{N\_shift\_mc}_B = 0.3/12 = 0.025$ and consequently, patient A is more complex than patient B.
5. M_shift_mth: mean of number of shifts (counted by the total number of visits to specialties) per month. Within a medical case, for each month the number of shifts (by the total number of visits to specialties) is calculated, next the mean is computed. The higher the mean, the higher the complexity of the medical case.

   Suppose that patients A and B have the sequences of visited specialties in the months January, February, March and April as shown in Table 2. Because $\text{M\_shift\_mth}_A = 0.3$ and $\text{M\_shift\_mth}_B = 0.2$, patient A is more complex than patient B.
6. Var_shift_mth: variance of number of shifts (counted by the total number of visits to specialties) per month. Within a medical case, for each month the number of shifts (counted by the total number of visits to specialties) is calculated, next the variance is computed. The higher the variance, the higher the complexity of the medical case. As we can see from Table 2, patient A is more complex than patient B.

Table 2
Example of visited specialties in 4 months (January, February, March and April) for patients A and B and the corresponding mean and variance

Patient	January	February	March	April	Mean	Variance
A	I-I-D	I-I-I	–	I-D-C	Mean (1/3, 0, 2/3) = 0.3	Var (1/3, 0, 2/3) = 0.11
B	I-I-I	C-I	I-I	I-I-D-I-I	Mean (0, 1/2, 0, 2/5) = 0.2	Var (0, 1/2, 0, 2/5) = 0.06

The six variables described above are used for developing logistically homogeneous groups within the population of patients with PAV diseases. If relevant clusters of patients can be found, these groups can be used in two ways: (i) to predict as early as possible to what cluster an individual patient belongs; (ii) to develop different logistic control systems for each homogeneous group. In this paper, we concentrate only on the first way of usage, namely to predict the cluster to which a patient is likely to be assigned. In the next section, we describe the clustering experiments in which we tried to find these logistically homogeneous groups. In Section 5, we try to develop predictive models based on the already developed logistically homogeneous groups.

3.3. Missing data

The existence of missing data should be carefully investigated in case of performing clustering analysis, because the possible missing data should be replaced with some estimates [5]. However, in our case we plan to cluster aggregated variables. The aggregation method chosen to operationalize the logistic complexity into logistic aggregate variables is filtering out (smoothing) missing values. Therefore, possible missing data that exist in the medical case log (our raw material) will not significantly affect the clustering results.

4. Development of logistic patient groups

After the selection and preparation of the data, our next step is the clustering of our patients with PAV diseases into, from the logistic point of view, homogenous groups that can be characterized by rules based on the aggregated logistic variables. In Section 5, we try to search for predictive rules that can be used to predict to which cluster new patients belong. First we turn to clustering.

4.1. Clustering experiments

Clustering techniques are used to group data into groups that are not known beforehand. As clustering method we chose two-step method, available in the Clementine 6.0.1 SPSS product [2]. The goal of this clustering technique is to (i) minimize variability within clusters; (ii) maximize the variability between clusters. The first step makes a single pass through the data, during which it compresses the raw input data into a manageable set of sub-clusters. The second step uses a hierarchical clustering method to progressively merge the sub-clusters into increasingly larger clusters, without requiring another pass through the data [2].

We chose this type of clustering technique because it shows two types of advantages. First, it is not necessary to decide beforehand the numbers of clusters. Second, compared to other techniques, it is faster for large datasets and a large number of variables. The technique seems therefore robust in case of a larger dataset and/or more variables.

For building the logistic patient groups, we ran two series of experiments: clustering experiments based on (i) all the six logistic variables built so far; (ii) factors extracted from the initial six logistic variables. For the later set of experiments, we use the factors

Table 3
Means and standard deviations for logistic variables in case of not clustered data ('total') and for the clustering model LOG_VAR_3 with three clusters

Logistic variables	Total	Clustering model LOG_VAR_3		
		Cluster-1 (2330[a])	Cluster-2 (127[a])	Cluster-3 (1938[a])
C_dif_visit				
Mean	3.51	2.566	3.976	4.608
S.D.	1.58	0.758	2.419	1.515
C_shift				
Mean	0.243	0.092	0.43	0.202
S.D.	0.217	0.139	0.215	0.138
N_visit_mc				
Mean	0.085	0.046	1.373	0.048
S.D.	0.286	0.074	0.997	0.045
N_shift_mc				
Mean	0.002	0.0	0.063	0.002
S.D.	0.026	0.002	0.142	0.004
M_shift_mth				
Mean	0.087	0.013	0.077	0.177
S.D.	0.113	0.025	0.13	0.112
Var_shift_mth				
Mean	0.029	0.005	0.006	0.06
S.D.	0.038	0.011	0.012	0.039

[a] Number of items in each cluster.

extracted with a principal component analysis technique, available also in the Clementine software.

4.2. Clustering experiment involving all logistic variables

In our first clustering experiment, all the logistic variables are used. We let the two-step method to search the number of clusters automatically. The results are given in Table 3.

The two-step method resulted in three clusters, with 2330, 127 and 1938 items. In order to choose the valid homogeneous clusters, we have to compare the standard deviation of each cluster with the standard deviation of the data not yet clustered. The cluster-1 and cluster-3 seem to show generally higher degrees of homogeneity compared with unclustered data. If we look for example in Table 3 at standard deviation values for variable C_dif_visit, both cluster-1 and cluster-3 have lower values than total ($0.758 < 1.58$ and $1.515 < 1.58$). In the following analyses, we therefore concentrate on cluster-1 and cluster-3.

We identified that it was possible to build two reliable clusters. But how can we interpret them? Different methods are available to characterize the clusters found by a clustering technique. One way to look at them is to investigate their means. However, in this paper we choose to use Quinlan's induction algorithm C4.5 rules [10] to characterize the clusters. Seven rules are induced to characterize cluster-1 and 12 rules for cluster-3. Examples of the induced rules are given in Table 4. For each rule, we have information about its coverage

Table 4
Some examples of the rules that characterize the different clusters based on all logistic variables

Rule number	Rule description	Coverage	Reliability (%)
Rule #1, cluster-1	IF C_dif_spm ≤ 3 and C_shift ≤ 0.296 and N_visit_mc ≤ 0.506 and M_shift_mth ≤ 0.101 and Var_shift_mth ≤ 0.042 THEN cluster-1	1943	99.9
Rule #11, cluster-3	IF C_dif_spm > 3 and C_shift > 0.304 and N_visit_mc ≤ 0.688 THEN cluster-3	1303	97.9

and the reliability. For instance, if we look at Rule #1 for cluster-1, there are 1943 examples covered by the IF-part of this rule, and 99.9% of them actually belong to cluster-1.

Inspecting the induced rules, the two clusters can be characterized as follows: cluster-1 includes ''moderately complex'' PAV patients, while cluster-3 covers the ''complex'' examples. As general characteristics, patients from the ''moderately complex'' cluster have visited up to three different specialists and show lower values for the shift characteristics, while patients from cluster ''complex'' have visited more than three different specialists and the values for shift features are higher. Cluster-2 seems to contain the 127 cases that cannot be grouped in cluster-1 or cluster-3. Two interesting rules (displayed in Table 5) are induced to characterize cluster-2.

The patients in cluster-2 show a higher number of visits counted by the duration of the medical case (variable N_visit_mc) than patients from cluster-1 and cluster-3, while the number of different specialists C_dif_spm is not so high. These rules give rise to the impression that patients who repeatedly visit one specialist are in this cluster. Inspection of the data reveals that these patients frequent the dialysis department. Because this is not a PAV-related cluster, we excluded this cluster from our further analysis.

4.3. Clustering experiment involving two latent factors

In the previous subsection, we applied the clustering technique directly to the six logistic variables. In this section, we first use a principal component analysis extraction method to

Table 5
Some examples of rules that characterize cluster-2 of the clustering model based on all logistic variables

Rule number	Rule description	Coverage	Reliability (%)
Rule #1, cluster-2	IF N_visit_mc > 0.688 THEN cluster-2	87	97.8
Rule #2, cluster-2	IF C_dif_spm ≤ 6 and N_visit_mc > 0.506 M_shift_mth > 0.074 THEN cluster-2	22	91.7

Table 6
Factor loadings for two latent factors extracted from the original six logistic variables

	Component	
	Factor-1	Factor-2
C_dif_spm	0.791	−0.027
C_shift	0.908	−0.056
N_visit_mc	0.044	0.890
N_shift_mc	0.056	0.894
M_shift_mth	0.848	0.035
Var_shift_mth	0.829	−0.084

check for possible latent factors. We then apply our clustering technique on these latent factors. Table 6 displays the results of the principal component analysis.

The total variance explained by this model is 74%. Inspecting the two extracted factors, the first factor can be observed showing high correlations with logistic variables C_shift, M_shift_mth, Var_shift_mth and C_dif_spm and very small correlations with the rest. The second factor show a high correlation with N_visits_mc and N_shift_mc and a low correlation with the other variables.

The factors are difficult to interpret; a hypothesis could be that these two factors represent two facets of complexity. Factor-1 represents somehow the ''complexity due to shifts'' and Factor-2 ''complexity in time span''. Thus, we can conclude that it is worthwhile to search for clusters based on these two factors. Table 7 shows the clustering model based on the extracted factors.

Similar to the previous experiments, by comparing the standard deviation of cluster-1 and cluster-3 with the standard deviation of data not yet clustered, cluster-1 and cluster-3 appear to have a higher homogeneity than the unclustered data. Just as a remark, the values of Factor-1 and Factor-2 for standard deviation and for the mean are 1 respectively 0 in case of unclustered data ('total' column in Table 7), because the principal component analysis extracts latent factors by standardizing the values of the input variables.

Again, we choose to use Quinlan's C4.5 rules induction algorithm to characterize the clusters. Twelve rules are found for cluster-1 and 16 for cluster-3, with confidences ranging

Table 7
Means and standard deviations for the two extracted latent factors, in case of not clustered data and in case of clustering model FACTOR_3 with three clusters

	Total	Clustering model FACTOR_3		
		cluster-1 (2936[a])	cluster-2 (154[a])	cluster-3 (1305[a])
Factor-1				
Mean	0	0.552	0.202	1.267
S.D.	1	0.547	0.81	0.565
Factor-2				
Mean	0	0.133	3.266	0.087
S.D.	1	0.136	4.11	0.163

[a] Number of items in each cluster.

Table 8
Some examples of the rules that characterize the different clusters based on two latent factors

Rule number	Rule description	Coverage	Reliability (%)
Rule #1, cluster-1	IF C_dif_spm ≤ 4 and C_shift ≤ 0.467 and N_visit_mc ≤ 0.492 and M_shift_mth ≤ 0.086 and Var_shift_mth ≤ 0.03 THEN cluster-1	2165	100
Rule #2, cluster-2	IF N_visit_mc > 0.604 THEN cluster-2	97	85.9
Rule #1, cluster-3	IF C_dif_spm > 4 and C_shift > 0.32 and Var_shift_mth > 0.027 THEN cluster-3	716	99.9

between 85 and 75%. Inspecting these rules (Table 8), we arrive at the similar conclusions: there is a cluster for ''moderately complex'' PAV patients and one for ''complex'' ones.

The rules look relatively similar, although there are some differences: (i) not surprisingly, more rules are based on factors; (ii) for each cluster, there is one rule with a very low coverage and also low confidence; we can interpret it as two rules which try to explain few cases which behave as exceptions. If we remove the two rules for ''exceptional'' cases for each cluster, we end up with 11 rules for cluster-1 and 15 rules for cluster-3, with confidence over 93 and 83%, respectively. Moreover, these clusters can be characterized by rules on which basis one cluster contains ''moderately complex'' PAV patients, and another one ''complex'' PAV patients, complexity being understood from the logistic point of view. The third cluster contains patients not especially suitable for our purposes: their logistic behaviour is determined only secondarily by PAV diseases.

The conclusion is that in both situations, (i) clustering based on all logistic variables; (ii) clustering based on two extracted logistic variables, we can obtain homogeneous logistic clusters. The question that we are trying to answer further is: can we use these clusters for prediction purposes? In the next section, we compare the two clusters for their capabilities to predict to which cluster a new individual patient belongs.

5. Development of predictive models

In the previous section, we saw that both clustering methods result in logistic homogeneous clusters. However, if it is not possible to predict to which cluster a new individual patient belongs, the clustering is of no use. In this section, we investigate if it is possible to use some a priori personal patient information such as age, gender and previous diagnoses, to predict what kind of logistic behaviour a patient newly entered in the process will have.

Apart from age and gender, a representation of the patient must be generated on the basis of his or her medical history, in order to be assigned to a particular cluster. Knowing to

which cluster a patient is likely to belong may provide immediate indications on how to plan future activities, capacity planning, etc.

In the following paragraphs, we describe how we develop predictive models that can be used to assign PAV patients to a certain logistic cluster, based on a priori information.

A priori information include age, gender, primary diagnosis, and potential secondary diagnoses. Age and gender are known for the first time when a patient is registered in the hospital and he/she receives a registration card. Primary diagnoses and potential secondary diagnoses are known only when the patient is clinically admitted. When a patient has a clinical admission, one mandatory primary diagnosis will be recorded and up to eight possible secondary diagnoses. For example, a patient can be admitted in the hospital because of acute gangrene as primary diagnosis; in the same time, this person has a chronic disease, namely arteriosclerosis as secondary diagnosis.

For developing predictive models, we use as learning material our database with 4395 medical cases, where the input attributes are age, gender and diagnosis. From the previous clustering phase, we already know for each record (i.e. medical case) to which cluster it belongs, thus each medical case is labeled as ''complex'' or ''moderately complex''. Note that it is possible for a patient to have one medical case that is ''moderate complex'' and another medical case that is ''complex''. In other words, the learning material is composed from records representing histories of medical cases rather than histories of patients.

Two series of learning experiments were performed for each clustering model, i.e. for the model based on all logistic variables LOG_VAR_3 and for the model based on two latent factors, FACTOR_3:

(i) Experiment ''*all diagnoses*'' with 60 input features: age, gender, total number of diagnoses and 57 possible diagnoses. Each diagnosis is represented as a separate binary feature; if a certain diagnosis is present in the medical case, the corresponding feature is marked with a ''1'' and with ''0'' if it is not present.
(ii) Experiment ''*chronic diagnoses*'' with 11 input features: age, gender, total number of diagnoses and eight diagnosis classes. For this experiment, we created eight diagnosis classes, in which we included all chronic diagnoses: (1) diabetes, (2) hypertension, (3) arteriosclerosis, (4) hyperhomocysteinemia, (5) hyperlipidaemia (including hypercholesterolaemia), (6) coagulation disorders, (7) heart problems and (8) (chronic) renal failure.

5.1. Experiment "all diagnoses"

In this first type of experiment, we consider 60 input features: age, gender, total number of diagnoses and 57 diagnoses. Each of the 57 diagnoses is taken as a separate feature. Here, we are interested to obtain predictive rules in which we can have combinations of age, gender, total number of diagnoses and individual diagnosis. The class (or output) feature is the cluster label, namely ''complex'' or ''moderately complex''. The experiment consists in training and afterwards testing the model, which will result in some rules of a certain quality. The training database contains the following fields:

Patient ID: number field.
Age: number field.

Gender: flag field (1 for male, 2 for female).
C_sec_diag: number field. Represents total number of diagnoses).
d***: flag field. This flag will be set to ''1'' if the patient has the diagnosis coded ''***'' within the medical case and to ''0'' if not. For example, if the patient has diabetes, which is coded ''250'', the feature d250 will be marked with ''1''.

5.2. Experiment ''chronic diagnoses''

This second type of experiment consists in 11 input features, namely age, gender, total number of diagnoses and eight groups of diagnoses. We consider the six chronic diagnoses (diabetes, hypertension, arteriosclerosis, hyperhomocysteinemia, hyperlipidaemia, coagulation disorders), heart problems and (chronic) renal failure, each one as separate features. Here we want to test whether specific chronic diseases and/or heart family of diseases can provide qualitatively predictive rules. Also, in this type of experiment, the class (or output) feature is the cluster label, namely ''complex'' or ''moderately complex''.

The database contains the following fields:

Patient ID: number field.
Age: number field.
Gender: flag field (1 for male, 2 for female).
C_sec_diag: number field. Represents total number of diagnoses.
g250, g401, g440, ...: flag fields. The diagnoses marked in these fields are all six chronic diagnoses. For example, g250 stands for diabetes, g401 for hypertension and g440 for arteriosclerosis. These flags will be set to ''1'' if the patient has within the medical case that specific diagnosis and to ''0'' if not.
hart: flag field. This flag will be set to ''1'' if the patient has within the medical case at least one diagnosis which relate to heart, and to ''0'' if not.
g585: flag field. This flag will be set to ''1'' if the patient has within the medical case the diagnosis coded 585 (renal failure), and to ''0'' if not.

We run in total four learning series: one experiment ''all diagnoses'' with clustering model LOG_VAR_3, one experiment ''all diagnoses'' with clustering model FACTOR_3, one experiment ''chronic diagnoses'' with clustering model LOG_VAR_3 and one experiment ''chronic diagnoses'' with clustering model FACTOR_3.

The quality of the predictive models is assessed by 10-fold cross-validation. This technique estimates the generalizing capacities of a learned model in the absence of a holdout test sample. Cross-validation is performed by dividing the training data into 10 subsets and then learning 10 models with each 10% subset held out in turn. The average accuracy of the models on the 10-holdout samples is used as an estimate of the accuracy of the model on new, ''unseen'' data. The cross-validation performance on test material for experiments ''all diagnoses'' and ''chronic disease'' with the two clustering models developed up to now, LOG_VAR_3 and FACTOR_3, is given in Table 9.

Because we want to compare the prediction performance of the models that we built so far, we repeat the development of other two alternative clustering models, one based on all logistic variables and one on the two extracted factors. We use the same two-step clustering

Table 9
Performance of predictive models from experiments ''all diagnoses'' and ''chronic diagnoses'', of clustering models based on all logistic variables (LOG_VAR_2 and LOG_VAR_3) and on two latent factors (FACTOR_2 and FACTOR_3)

Model	No. of elements in each cluster	No. of clusters	Baseline performance	All diagnoses		Chronic diagnosis	
				Performance	Gain	Performance	Gain
LOG_VAR_2	cluster-1: 2330 (53.01%); cluster-2: 2065 (46.99%)	2	53	61.2	8.2	63.3	10.3
FACTOR_2	cluster-1: 2936 (66.80%); cluster-2: 1459 (33.20%)	2	67	68.5	1.5	69.4	2.7
LOG_VAR_3	cluster-1: 2330 (53.01%); cluster-2: 127 (2.89%); cluster-3: 1938 (44.10%)	3	53	58.6	5.6	60.5	7.5
FACTOR_3	cluster-1: 2936 (66.80%); cluster-2: 154 (3.51%); cluster-3: 1305 (29.69%)	3	67	64.1	–	64.6	–

method, but we do not let the method find the number of clusters automatically. Rather, we fix the number of final clusters at 2.

The resulting model LOG_VAR_2 consists of two clusters: cluster-1 that contains the same cases (2330) like the ''moderately complex'' cluster from clustering model LOG_VAR_3, and cluster-2 which joins the rest of the cases (2065). In the same manner, model FACTOR_2 yields two clusters: cluster-1 that contains the same cases (2936) as cluster ''moderately complex'' from clustering model FACTOR_3, and cluster-2 which joins the rest of the cases (1459). The performance of these two models is also shown in Table 9.

Of interest are models that show a higher performance than the *baseline performance* (the percentage of the most common class; in our case, in model LOG_VAR_2, cluster-1 comprises 53% of all elements; if the model always predict cluster-1, a performance level of 53% would be attained). As can be seen from Table 9, the predictive model with the highest gain in performance concerns the experiment with ''chronic diagnoses'', where the cases are labeled based on clusters developed with model LOG_VAR_2 (all logistic variables and two clusters). Its overall performance is 63; 10% higher than baseline class guessing. The predictive models based on clustering models LOG_VAR_2 and LOG_-VAR_3 also show a certain gain over the baseline performance. In contrast, the clusters based on the two latent factors show very small gain over the baseline performance, if any.

To illustrate what is learned, we concentrate on the rules of the predictive models from experiment ''chronic diagnoses'' in case of LOG_VAR_3. They are presented in Table 10.

The five rules developed for cluster-1 can be shared in two categories: the first three, Rule #1, Rule #2 and Rule #3, which show a low support (5, 16 and 6 respectively) and a high confidence (85.7, 83.3 and 75%) and Rule #4 and Rule #5, with a high support (2197 and 3098) and low confidence (62.4 and 61.1%). Because we are interested not only in having high performance (rules with high confidence), but certainty also in wide-coverage general rules that may provide new useful knowledge, we inspect rules Rule #4 and Rule #5

Table 10
Predictive rules from experiment ''chronic diagnoses'' with clustering model LOG_VAR_3

Rule number	Rule description	Coverage	Reliability (%)
Rule #1, cluster-1	IF C_sec_diag > 2 and C_sec_diag ≤ 3 and g250 = F and g272 = T THEN cluster-1	5	85.7
Rule #2, cluster-1	IF Age > 80 and C_sec_diag ≤ 3 and g401 = T THEN cluster-1	16	83.3
Rule #3, cluster-1	IF Age > 91 and C_sec_diag > 2 and C_sec_diag ≤ 3 THEN cluster-1	6	75.0
Rule #4, cluster-1	IF Age ≤ 72 and C_sec_diag ≤ 3 and g250 = F and g585 = F THEN cluster-1	2197	62.4
Rule #5, cluster-1	IF C_sec_diag ≤ 2 and g585 = F THEN cluster-1	3098	61.1
Rule #1, cluster-3	IF Age > 65 and Age ≤ 68 and C_sec_diag > 2 and C_sec_diag ≤ 3 and g272 = F and g401 = T and hart = F THEN cluster-3	11	92.3
Rule #2, cluster-3	IF g585 = T THEN cluster-3	93	65.3
Rule #3, cluster-3	IF Age ≤ 72 and Gender = 2 and C_sec_diag > 2 and C_sec_diag ≤ 3 and g272 = F and g401 = F and hart = F and g585 = F THEN cluster-3	53	61.8
Rule #4, cluster-3	IF C_sec_diag > 2 THEN cluster-3	1259	60.1

more closely. Using the same reasoning for the rules induced to capture cluster-3, we focus on Rule #2, Rule #3 and Rule #4.

We recall that this experiment type ''chronic diagnoses'' focus on a priori characteristics, i.e. age, gender, total number of diagnoses and eight groups of diagnoses: diabetes (g250), hypertension (g401), arteriosclerosis (g440), hyperhomocysteinemia (g2704), hyperlipidaemia (g272), coagulation disorders (g286), heart problems (hart) and (chronic) renal failure (g585).

The wide-coverage rules tell us that if a patient has three or less diagnoses, and does not have diagnosis g585 (renal failure), it is likely that he/she will be in cluster-1: a ''moderately complex'' patient. In contrast, if a patient has diagnosis g585 (renal failure), it will be a ''complex'' patient. Also, according to Rule #4 for cluster-3, if the number of diagnoses is higher than 2, it will estimated to be a ''complex'' patient. If the patient does

Table 11
Predictive rules from experiment "chronic diagnoses" with clustering model LOG_VAR_3

Rule number	Rule description	Coverage	Reliability (%)
Rule #1, cluster-1	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag ≤ 2 and d286 = 0 and d250 = 1 and d7802 = 0 and d440 = 0 and d4359 = 0 and d2508 = 0 and Age > 55 THEN cluster-1	98	61.2
Rule #5, cluster-1	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag ≤ 2 and d286 = 0 and d250 = 0 and d425 = 0 and d997 = 0 and d446 = 0 and d413 = 0 and d428 = 0 and d426 = 0 and d441 = 0 and d443 = 0 and d707 = 0 and d2508 = 0 THEN cluster-1	2383	63.8
Rule #7, cluster-1	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag > 2 and C_sec_diag ≤ 5 and d447 = 0 and d2508 = 0 and d443 = 0 and d403 = 0 and d437 = 0 and d446 = 0 and d9972 = 0 and d357 = 0 and d250 = 0 and d426 = 0 and d410 = 1 and d4331 = 0 and d436 = 0 and d707 = 0 and d413 = 0 and d7854 = 0 and d997 = 0 and d412 = 0 and d998 = 0 THEN cluster-1	51	68.6
Rule #9, cluster-1	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag > 2 and d447 = 0 and d2508 = 0 and d443 = 0 and d403 = 0 and d437 = 0 and d446 = 0 and d9972 = 0 and d357 = 0 and d250 = 0 and d426 = 0 and d410 = 0 and d427 = 0 and d459 = 0 and d5571 = 0 and d442 = 0 and d997 = 0 and d424 = 0 and d425 = 0 and d428 = 0 and d4359 = 0 and d2720 = 0 and d413 = 0 and d412 = 0 and d444 = 0	56	62.5

Table 11 (*Continued*)

Rule number	Rule description	Coverage	Reliability (%)
	THEN cluster-1		
Rule #1, cluster-3	IF d585 = 1 and d442 = 0 and d429 = 0 and d444 = 0 THEN cluster-3	74	70.3
Rule #2, cluster-3	IF d585 = 0 and d2507 = 1 and C_sec_diag ≤ 7 and d414 = 0 and d250 = 1 THEN cluster-3	53	79.2
Rule #8, cluster-3	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag > 2 and d447 = 0 and d2508 = 1 THEN cluster-3	47	78.7
Rule #13, cluster-3	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag > 3 and d447 = 0 and d2508 = 0 and d443 = 0 and d403 = 0 and d437 = 0 and d446 = 0 and d9972 = 0 and d357 = 0 and d250 = 0 and d426 = 0 and d410 = 0 and d427 = 1 and d4331 = 0 THEN cluster-3	47	91.5
Rule #14, cluster-3	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag > 2 and d447 = 0 and d2508 = 0 and d443 = 0 and d403 = 0 and d437 = 0 and d446 = 0 and d9972 = 0 and d357 = 0 and d250 = 0 and d426 = 0 and d410 = 0 and d427 = 0 and d459 = 0 and d5571 = 0 and d442 = 0 and d997 = 1 and d412 = 0 THEN cluster-3	66	62.1

not have diagnosis g585, g401, g272 and "hart" (heart) problems, has in total three diagnoses and is a woman, then she has some chance to be a "complex" patient.

However, the rules provided by this predictive model provide restricted information, regarding only the chronic and heart problems. What if a patient does not suffer from such diseases? More detailed rules, at the level of individual diagnoses are provided by the predictive model from experiment "all diagnoses" with LOG_VAR_3. A selection of the rules with support higher than 20 instances and confidence higher than 0.6 is shown in Table 11.

Among the rules induced for cluster-1, we inspect Rule #1: if a patient has diagnosis d250 (diabetes) (and does not have the other eight specified diagnoses), has two or less than

two diagnoses and age more than 55, he or she is likely to be a ''moderately complex'' patient. Looking at Rule #2 for cluster-3, we can notice that if a patient has in addition to diagnosis d250 the diagnosis d2507 (diabetic foot), it will be assigned to cluster-3, which is the cluster for ''complex'' patients. Subsequently, the number of diagnoses will be higher, which is also according to the rule (number of diagnoses C_sec_diag ≤ 7). Thus, our model contains a rule that is able to ''send'' the patient to the right cluster, when an additional diagnosis becomes known.

Another meaningful rule is Rule #1 for cluster-3, which says that if a patient has diagnosis d585 (renal failure) and do not have the other three specified diagnoses, he/she will be a ''complex'' patient. Thus, this rule provides a way to distinguish the patients who need dialysis and it can be expected that they will be ''complex'' patients.

6. Discussion

Our first goal was to see whether patients with PAV disease could be clustered into logistically homogeneous groups. The two different clustering models that we developed, both based on all six logistic variables and two latent factors, show that some reliable clustering is possible. This result can be used as a starting point for building alternative classification models that look for homogeneity from the logistic point of view and not only from the medical point of view.

The two considered approaches, i.e. clustering on logistic variables, and clustering based on latent factor extracted from logistic variables, both lead to three main clusters, of which two hold clear-cut groups of patients: one can be labeled ''moderately complex'' patients, while the other holds ''complex'' patients. The remaining third cluster contains a small number of cases that cannot be assimilated to one of the two valid clusters. The rules induced for the characterization of each cluster provide a good insight into the relative importance of the involved logistic dimensions, and here we recall them: (1) C_dif_spm, (2) C_shift, (3) N_visit_mc, (4) M_shift_mth and (5) Var_shift_mth, all these computed per medical case. The rules indicate, for instance, that N_shift_mc may have a low importance: it is never used in any of the rules in the rule set. Tests based on this feature are removed from the rules because they do not contribute enough, apparently, to the classification power of the model. Next to providing information about the logistic variables, the induced rules that distinguish between ''complex'' patients and ''moderately complex'' patients can eventually provide reasons for developing a control system.

The grouping models that we develop are fully useful if we are able to combine them with predictive models. Therefore, we are interested to develop predictive models that uses a priori information to predict in which cluster a patient is likely to be, as soon as the patient enters the healthcare system. The predictive models obtained so far are rather general. Nevertheless, we can extract some useful information. Look for example to the following rules produced in experiment ''all diagnoses'' with clustering model LOG_VAR_3, shown in Table 12.

Rule #1 for cluster-1 says that a patient is ''moderately complex'' if he/she does not have diagnosis d585 (renal failure), d2507 (diabetic foot), d429, d286, but has d250 (diabetes) and C_sec_diag ≤ 2. In contrast, using Rule #2 for cluster-3, a patient is estimated to be

Table 12
A selection of predictive rules from experiment "all diagnoses" with clustering model LOG_VAR_3

Rule number	Rule description	Coverage	Reliability (%)
Rule #1, cluster-1	IF d585 = 0 and d2507 = 0 and d429 = 0 and C_sec_diag ≤ 2 and d286 = 0 and d250 = 1 and d7802 = and d440 = 0 and d4359 = 0 and d2508 = 0 and Age > 55 THEN cluster-1	98	61.2
Rule #1, cluster-3	IF d585 = 1 and d442 = 0 and d429 = 0 and d444 = 0 THEN cluster-3	74	70.3
Rule #2, cluster-3	IF d585 = 0 and d2507 = 1 and C_sec_diag ≤ 7 and d414 = 0 and d250 = 1 THEN cluster-3	53	79.2

"complex" if he/she additionally has diagnosis d2507 (and not diagnosis d585 and d414), increasing the number of diagnoses, i.e. C_sec_diag $\leq$ 7. Rule #1 for cluster-3 expresses that as soon as a patient has diagnosis d585 (renal failure), it will be a complex patient (a PAV patient that need dialysis as well). It should be noted that the models presented here are based on a relatively small set of examples, and their outcomes should be taken as indicative of their potential; until there is considerably more data, the obtained predictive rules are not detailed enough and reliable to base a whole control system on.

7. Conclusions and future work

In the present paper, we proposed a methodology that attempts to offer a solution for a better coordination of patients with peripheral vascular diseases. We showed that by using clustering technique and factor analysis, PAV patients can be shared in two clear-cut clusters, namely "complex" and "moderately complex" patients. These clustering models are relevant if predictive models can be built, based on some a priori patient characteristics. Using data mining techniques, we developed such predictive models and we illustrated that rules can found. The rules that assign patients to clusters also provide clues about which of the six logistic variables that represent a medical case are relevant or not, and in which interaction they are relevant.

Further research should be invested in finding more a priori patient characteristics that allow predicting logistic clusters more reliably. We plan to do future research by developing a multi-step model. A priori knowledge as age, gender, risk factors and relevant secondary diagnosis are known the first time a patient enters the hospital. Based on these information, a first prediction could be made and patients could receive the proper treatment faster. Also, when more information become available through time (as more

steps in the process become known), a secondary more precise prediction can be made. Thus, changes in patient groups and treatments could automatically be discovered and relayed back to the hospital management to inspect whether the new data warrant new changes.

References

[1] Bertrand JWM, Wortmann JC, Wijngaard J. Production control. A structural and design oriented approach. Amsterdam: Elsevier, 1990.
[2] Clementine Datamining System, version 6.0.1. User guide. SPSS Inc., 2000.
[3] CaseMix Quarterly of the Patient Classification System Europe organization Web Site. Available at http://www.casemix.org.
[4] Cios KJ, Teresinka A, Konieczna S, Potocka J, Sharma S. Diagnosing myocardial perfusion SPECT bull's-eye maps: a knowledge discovery approach. IEEE Eng Med Biol 2000;19(4):17–25.
[5] Dilts D, Khamalah J, Plotkin A. Using clustering analysis for medical resource decision making. Med Decision Mak 1995;15(4):333–47.
[6] Fetter RB. The new ICD-9-CM Diagnosis-Related Group classification scheme, HCFA Publication no. 03167. Washington: Health Care Financing Administration, US Government Printing Office, 1983.
[7] Fetter RB, Averill A. Ambulatory visit groups: a framework for measuring the productivity in ambulatory care. Health Serv Res 1984;19:415–37.
[8] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89–109.
[9] Lavrač N. Selected techniques for data mining in medicine. Artif Intell Med 1993;16:3–23.
[10] Quinlan JR. C4.5: Programs for machine learning. Los Altos (CA): Kaufmann (Morgan), 1993.
[11] Miksch S. Plan management in the medical domain. AI Commun 1999;12:209–35.
[12] Ploman M. Choosing a patient classification system to describe the hospital product. Hosp Health Serv Admin 1985;(5/6):106–17.
[13] Spyropoulos CD. AI planning and scheduling in the medical hospital environment. Artif Intell Med 2000;20:101–11.
[14] Vissers JMH. Patient flow-based allocation of hospital resources, Doctoral Thesis. The Netherlands: Eindhoven University Press, 1994.
[15] de Vries G, Bertrand JWM, Vissers JMH. Design requirements for health care production control systems. Prod Plann Control 1999;10(6):559–69.
[16] de Vries GG, Vissers JMH, de Vries G. The use of patient classification systems for production control of hospitals. Casemix Quart 2000;2(2):65–70.
[17] de Vries GG, Vissers JMH, de Vries G. Logistic control system for medical multi-disciplinary patient flows. In: De Angelis V, Ricciardi N, Storchi G, editors. Monitoring, evaluating, planning health services. Proceedings of the 24th Meeting of the Working Group on Operational Research Applied to Health Services, ORAHS '98, 1998 July 19–24; Singapore: World Scientific, 1999. p. 141–51.