# 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] Technical University Eindhoven, the Netherlands
[b] Tilburg University, the Netherlands
[c] University of Antwerp, Belgium
[d] Prismant Institute for Health Care Management, Utrecht, the Netherlands

Present-day health care witnesses a growing demand for coordination of patient care. Coordination is needed especially in those cases in which hospitals have structured health care into specialty-oriented units, while a substantial portion of patient care is not limited to single units. From a logistical point of view, this multidisplinary patient care creates a tension between on the one hand controlling the hospital's units, and on the other hand the need for a control of the patient flow among 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 salient patients 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 medical 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 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 multidisciplinary units, and (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 peripheral arterial vascular diseases. Two groupings and the accompanied classification rule sets are presented. One grouping is based on all logistic variables, and another grouping is based on two variables, found by applying factor analysis. On the basis of the results we can conclude that the search for logistic homogenous groups has advantages over more traditional search for medically homogenous groups.

## 1. Introduction

In the Netherlands, as in many 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 health care expenditures. It is a recognised problem that suboptimally 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 health care into specialty-oriented units, and care for patients is not constrained within single units. From a logistical point of view this creates a tension between on the one hand the control over the units, and on the other hand the coordination needed among units to control the patient flow.

The total flow of patients in a hospital can be divided into mono-disciplinary patients 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. 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 (DRG) [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]. Thus, a solution will be to consider a logistic classification system that can assure a higher 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 were successfully used 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 threefold 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 will operationalise the concept of logistic complexity into different aggregate logistic variables that will be used further in clustering. Subsequently, we will characterize the obtained clusters by rules based on the aggregated logistic variables.
(ii) Second, we aim at developing a rule predictive model which 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 as soon as possible the new patient to a cluster.
(iii) Third, we illustrate how data mining techniques can help in this process.

We plan to assess the quality of the logistic clusters considering a combination of different criteria. In the first row, we want that our obtained clusters to be more ***homogeneous*** than the data not yet clustered. In the second row, 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 detailed 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 logistical 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 logistical complexity concept. Section 4 describes the clustering experiments for

finding logistical 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 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. 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 brains) 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 health care to know what specialty to refer to; knowledge within the hospital is dispersed; there is a lack of within-hospital co-operation; 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 pathologies.

**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<br>thrombus | ischaemia | pain in legs | impair of organs, muscles and arteries |
| disturbed composition of the blood | plaque<br>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 |
| etc. | etc. | etc. | etc. | etc. |

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. Before translating medical goals into data mining goals we need an extra step of logistical goals.

### 2.2 Logistical problem domain

Before we can elaborate on data mining goals, we have to expound our view on logistics. 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 health care 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 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. What we are looking for from a logistical point of view is 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 logistic homogeneous groups is important, because every logistic group can require its own optimal control system.

## 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 the complexity of the care process and the routing variability of the care process. In this paper we will concentrate on the complexity of the care process of a patient. Keep in mind we are referring to logistical complexity, which can be something completely different as medical complexity. Before we come to the part of operationalisation of 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, out-patient databases, laboratory databases, etc. When all patient information is gathered we have the hospital history of the patient.

We plan to do cluster analysis in order to find the logistic homogeneous groups. In order to perform cluster analysis, we have to assure that "some critical steps should be followed" [5]. These steps involve the selection of the (i) sample (population), (ii) attribute (or variables) that should be recorded (measured), (iii) similarity measure used to compare the entities, and the choice of the (iv) clustering technique(s).

### 3.1 Data selection

The first step for data sampling was to establish the criteria based on which a patient is a PAV patient. For this purpose, interviews were held with specialists from the source hospitals, which resulted in three diagnosis lists: (a) degenerative underlying chronic diseases, (b) PAV diseases and (c) diagnosis related with chronic or PAV diseases. We selected all patients who have at least one diagnosis from list (a) or (b) from the Elisabeth Hospital located in Tilburg, the Netherlands. For these patients, all records related to visits in different departments of the hospital were extracted. These records contain information mainly related with:

- personal characteristics: age, gender, address, date of birth and date of death if is the case, etc.
- characteristics of the policlinic visit: specialist, date, referral date, referring specialist (if the general practitioner request the visit or another specialist from the hospital), urgency, etc.
- characteristics of the clinical admission: specialist, date, diagnosis (1 main diagnosis and up to 8 possible secondary diagnosis), treatment, referring specialist (if the general practitioner request the admission or another specialist from the hospital), urgency or planned admission, etc.
- radiology, functional investigations information, other investigations.

These information were used to build a time-ordered history for 3603 patients. Please note that our purpose is not to analyze the underling processes in the patient's history. For instance, given a patient who breaks a leg in February, and undergoes an appendectomy in August, we will 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 to be used in clustering.

## 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 following:

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: Internal medicine – Internal medicine – Cardiology – Internal medicine – Dermatology – Internal medicine. Thus, the logistic variable C_dif_visit = 3.
2. **C_shift**: number of shifts within the medical case, accounted 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, accounted by the total number of visits, the more complex the medical case. As an illustration, let's have a look to the following example, in which "I" represents Internal medicine, "C"- cardiology and "D" – dermatology. 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. $C\_shift_A = 4/6 = 0.6$. 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, $C\_shift_B = 4/13=0.3$. 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, accounted 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 four weeks, whether patient B visited three specialties in twelve weeks. Subsequently, $N\_visit\_mc_A = 3/4 = 0.7$ and $N\_visit\_mc_B = 3/12 = 0.2$, consequently patient A is more complex than patient B.
4. **N_shift_mc**: number of shifts within the medical case per time-scale, accounted 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 four weeks: I-I-C-I-D-I. Patient B visited the following specialties in twelve weeks: I-I-C-I-I-I-I-D-I-I-I-I-I. Hence, $N\_shift\_mc_A = 0.6/4 = 0.15$, $N\_shift\_mc_B = 0.3/12=0.025$ and consequently, patient A is more complex than patient B.

5. **M_shift_mth**: mean of number of shifts (accounted by the total number of visits to specialties) per month. Within a medical case, for each month the number of shifts (accounted 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 $M\_shift\_mth_A = 0.3$ and $M\_shift\_mth_B = 0.2$, patient A is more complex than patient B.
6. **Var_shift_mth**: variance of number of shifts (accounted by the total number of visits to specialties) per month. Within a medical case, for each month the number of shifts (accounted 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 four months (January, February, March and April) for patient A and B and the correspondent 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 |

These six variables described above will be used further for developing logistic 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 patients belongs and (ii) to develop different logistic control systems for each homogeneous group. In this paper, we will concentrate only on the first way of usage, namely how can be used the developed clusters for prediction purposes. We describe in the next section the clustering experiments in which we tried to find these logistic homogeneous groups. In Section 5, we will try to develop predictive models based on the already developed logistic homogeneous groups.

# 4. Development of logistic patient groups

The question that we want to answer is: can our patients with PAV diseases be distinguished into meaningful clusters, from the logistical point of view?

## 4.1 Clustering experiments

Clustering techniques are used to group data into groups which are not known a-priori. As clustering method we chose Two-Step method, available in the Clementine 6.0.1 SPSS product. The goal of this clustering technique is to (i) minimize variability within clusters and (ii) maximize 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 larger and larger clusters, without requiring another pass through the data [2].

We chose this type of clustering technique because it shows two types of advantages: (i) it is not necessary to decide beforehand the numbers of clusters; (ii) compared to other techniques, it is faster for

large data sets because of its initial pre-clustering technique.

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

## 4.2 Clustering experiment involving all logistic variables

The first clustering experiment involves all logistic variables and search for the best numbers of clusters. The results are given below in Table 3.

**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.**

| | | **Total** | Clustering model **LOG_VAR_3** | | |
|---|---|---|---|---|---|
| | | | ***cluster-1*** | *cluster-2* | ***cluster-3*** |
| No. of items in clusters / Logistic variables | | - | 2330 | 127 | 1938 |
| C_dif_visit | Mean | 3.51 | 2.566 | 3.976 | 4.608 |
| | Std. Deviation | 1.58 | 0.758 | 2.419 | 1.515 |
| C_shift | Mean | 0.243 | 0.092 | 0.43 | 0.202 |
| | Std. Deviation | 0.217 | 0.139 | 0.215 | 0.138 |
| N_visit_mc | Mean | 0.085 | 0.046 | 1.373 | 0.048 |
| | Std. Deviation | 0.286 | 0.074 | 0.997 | 0.045 |
| N_shift_mc | Mean | 0.002 | 0.0 | 0.063 | 0.002 |
| | Std. Deviation | 0.026 | 0.002 | 0.142 | 0.004 |
| M_shift_mth | Mean | 0.087 | 0.013 | 0.077 | 0.177 |
| | Std. Deviation | 0.113 | 0.025 | 0.13 | 0.112 |
| Var_shift_mth | Mean | 0.029 | 0.005 | 0.006 | 0.06 |
| | Std. Deviation | 0.038 | 0.011 | 0.012 | 0.039 |

Comparing the standard deviations for each cluster with the standard deviation for the data not yet clustered, we can check the quality of the cluster homogeneity. Thus, cluster-1 and cluster-3 seem to show higher degrees of homogeneity compared with unclustered data. In the following analyses we will therefore concentrate on cluster-1 and cluster-3.

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 rule induction algorithm [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 Figure 1. The information between round brackets in the THEN-part of the rules indicates the covering and the reliability of the rule. For instance, (1943, 0.999) after the first rule of cluster-1 indicates that 1943 examples are covered by the IF-part of this rule, and 0.999 of them actually belong to cluster-1.

**Figure 1: Some examples of the rules that characterize the different clusters based on all logistic variables.**

```
Rule #1 for 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, 0.999)
Rule #11 for cluster-3:
  if  C_dif_spm > 3 and  C_shift > 0.304 and  N_visit_mc <= 0.688
  then -> cluster-3 (1303, 0.979)
```

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 Figure 2) are induced to characterize this rest cluster.

**Figure 2: Some examples of rules that characterize cluster-2 of the clustering model based on all logistic variables.**

```
Rules for cluster-2:
  Rule #1 for cluster-2:
   if  N_visit_mc > 0.688 then -> cluster-2 (87, 0.978)

Rule #2 for cluster-2:
   if  C_dif_spm <= 6 and  N_visit_mc > 0.506 and  M_shift_mth > 0.074
   then -> cluster-2 (22, 0.917)
```

The patients in cluster-2 show a higher number of visits accounted by the duration of the medical case (variable N_visit_mc) than patients from cluster-1and 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 on the six logistic variables. In this section we first use a Principal Component Analysis extraction method to check for possible latent factors. We then apply our clustering technique on these latent factors. Table 4 displays the results of the Principal Component Analysis.

**Table 4. Factor loadings for two latent factors extracted from the original six logistic variables.**

**Component Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| C_dif_spm | ,791 | -2,77E-02 |
| C_shift | ,908 | 5,613E-02 |
| N_visit_mc | -4,42E-02 | ,890 |
| N_shift_mc | 5,659E-02 | ,894 |
| M_shift_mth | ,848 | 3,502E-02 |
| Var_shift_mth | ,829 | -8,44E-02 |

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

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 clusters based on these two factors. Table 5 shows the clustering model based on extracted factors.

**Table 5. 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*** | *cluster-2* | ***cluster-3*** |
| Number of items in clusters | | - | 2936 | 154 | 1305 |
| Factor-1 | Mean | 0 | 0.552 | 0.202 | 1.267 |
| | Std. Deviation | 1 | 0.547 | 0.81 | 0.565 |
| Factor-2 | Mean | 0 | 0.133 | 3.266 | 0.087 |
| | Std. Deviation | 1 | 0.136 | 4.11 | 0.163 |

Similar to the previous experiments, cluster-1 and cluster-3 appear to have a higher homogeneity than the unclustered data. Again, we choose to use Quinlan's rule induction algorithm to characterize the clusters. 12 rules are found for cluster-1 and 16 for cluster-3, with confidences ranging between 0.85 and 0.75. Inspecting these rules (Figure 3), we arrive at the similar conclusions: there is a cluster for "moderately complex" PAV patients and one for "complex" ones.

**Figure 3: Some examples of the rules that characterize the different clusters based on two latent factors.**

```
Rules for cluster-1:
 Rule #1 for 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, 1.0)

Rules for cluster-2:
  Rule #4 for cluster-2:
   if  N_visit_mc > 0.604 then -> cluster-2 (97, 0.859)

Rules for cluster-3:
  Rule #1 for cluster-3:
   if  C_dif_spm > 4 and  C_shift > 0.32 and  Var_shift_mth > 0.027
   then -> cluster-3 (716, 0.999)
```

The rules look relatively similar, although there are some differences: (i) not surprisingly, more rules are based on factors and (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 0.93 and 0.83, respectively. One cluster contains "moderately complex" PAV patients, and another one "complex" PAV patients, complexity being understood from the logistical point of view. The third cluster contains patients not especially suitable for our purposes: their logistical behavior is determined only secondarily by PAV diseases.

A general observation is that in both situations, (i) clustering based on all logistical variables and (ii) clustering based on two extracted logistical variables, we can obtain homogeneous clusters. 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 diagnosis, to predict what kind of logistic behavior 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, 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 receive a registration card. Primary diagnoses and potential secondary diagnosis are known only when the patient is clinically admitted. When a patient has a clinical admission, it will be recorded one mandatory primary diagnosis and up to eight possible secondary diagnosis. 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 will use as learning material our database with 4395 medical cases, where the input attributes are age, gender and diagnosis. We already know each record (i.e. medical case) to which cluster belongs from the previous clustering phase, thus the cluster label will be the output. Just as a remark, it is perfectly possible that a patient can have one medical case that belongs to, let's say, "moderate complex" and another medical case which belongs to cluster "complex". Therefore, the learning material is composed from records represented as histories of medical cases, and not as histories of patients.

Two series of 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 diagnosis"*** has 60 input features: age, gender, total number of diagnosis and 57 possible diagnosis. Each diagnosis is represented as a separate 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 diseases"*** has 11 input features: age, gender, total number of diagnosis and 8 diagnosis classes.
For this experiment, we created 8 diagnosis classes, in which we included all chronic diseases: diabetes, hypertension, arteriosclerosis, hyperhomocysteinemia, hyperlipidaemia (including hypercholesterolaemia), coagulation disorders, heart problems and (chronic) renal failure.

## Experiment "all diagnosis"

Given a-priori knowledge as age, gender, total number of different diagnosis and 57 diagnosis, predict the cluster. In this experiment, each diagnosis is taken as a separate feature. 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 diagnosis. |
| ***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. |

## Experiment "chronic diagnosis"

Given a-priori knowledge as age, gender, total number of diagnosis and 8 groups of diagnosis, predict the cluster. In this experiment, we consider all 6 chronic diseases as separate features, plus heart problems and (chronic) renal failure. 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 diagnosis.
***g250***, ***g401***, ***g440***, ... : flag fields. The diagnosis marked in these fields are all 6 chronic diagnosis (for example, g250, g401 and g440 stands for diabetes, hypertension and arteriosclerosis, respectively). These flags will be set to "1" if the patient has that specific diagnosis within the medical case and to "0" if not.
***hart***: flag field. This flag will be set to "1" if the patient has at least one diagnosis which relate to heart, within the medical case and to "0" if not.
***g585***: flag field. This flag will be set to "1" if the patient has diagnosis coded 585 (renal failure), within the medical case and to "0" if not.

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 hold-out test sample. Cross-validation is performed by dividing the training data into ten subsets and then learning ten models with each 10% subset held out in turn. The average accuracy of the models on the ten hold-out samples is used as an estimate of the accuracy of the model on new data, in our case, new medical case. The cross-validation performance on test material for experiments "all diagnosis" and "chronic disease" with the two clustering models developed up to now, LOG_VAR_3 and FACTOR_3, is given below in Table 6.

For reasons of comparison, we repeated the development of two alternative clustering models, based on all logistic variables and on the two extracted factors. We used the same Two Step clustering method, but we fixed the number of final clusters to be 2. The resulting model LOG_VAR_2 consists on 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 6.

**Table 6. Performance of predictive models from experiments "all diseases" and "chronic diseases", based on all logistic variables (LOG_VAR_2 and LOG_VAR_3) and on two latent factors (FACTOR_2 and FACTOR_3).**

| Model | Number of elements in each cluster | Number of clusters | Baseline perfor-mance | experiment "all diagnosis" | | experiment "chronic diseases" | |
|---|---|---|---|---|---|---|---|
| | | | | *perf* | *gain* | *perf* | *gain* |
| **LOG_VAR_2** | cl-1: 2330 53.01%<br>cl-2: 2065 46.99% | 2 | *53* | 61.2 | **8.2** | 63.3 | **10.3** |
| **FACTOR_2** | cl-1: 2936 66.80%<br>cl-2: 1459 33.19% | 2 | *67* | 68.5 | 1.5 | 69.4 | 2.7 |
| **LOG_VAR_3** | cl-1: 2330 53.01%<br>cl-2: 127 2.88%<br>cl-3: 1938 44.09% | 3 | *53* | 58.6 | **5.6** | 60.5 | **7.5** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **FACTOR_3** | cl-1: 2936 66.80%<br>cl-2: 154 3.50%<br>cl-3: 1305 29.69% | 3 | *67* | 64.1 | - | 64.6 | - |

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 6, the predictive model with the highest gain in performance concerns the experiment with "chronic diseases", where the cases are labeled based on clusters developed with model LOG_VAR_2 (all logistic variables and 2 clusters). Its overall performance is 63%; 10% higher than baseline class guessing. The other three 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 diseases" in case of LOG_VAR_3. They are presented below (Figure 4).

**Figure 4: Predictive rules from experiment "chronic diseases" with clustering model LOG_VAR_3.**

```
Rules for cluster-1:
    Rule #1 for cluster-1:
        if  C_sec_diag > 2 and  C_sec_diag <= 3 and  g250 == F and  g272 == T
          then -> cluster-1 (5, 0.857)
    Rule #2 for cluster-1:
        if  Age > 80 and  C_sec_diag <= 3 and  g401 == T then -> cluster-1 (16, 0.833)
    Rule #3 for cluster-1:
        if  Age > 91 and  C_sec_diag > 2 and  C_sec_diag <= 3 then -> cluster-1 (6, 0.75)
    Rule #4 for cluster-1:
        if  Age <= 72 and  C_sec_diag <= 3 and  g250 == F and  g585 == F
          then -> cluster-1 (2197, 0.624)
    Rule #5 for cluster-1:
        if  C_sec_diag <= 2 and  g585 == F then -> cluster-1 (3098, 0.611)

Rules for cluster-3:
    Rule #1 for 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, 0.923)
    Rule #2 for cluster-3:
        if  g585 == T then -> cluster-3 (93, 0.653)
    Rule #3 for 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, 0.618)
    Rule #4 for cluster-3:
        if  C_sec_diag > 2 then -> cluster-3 (1259, 0.601)
Default : -> cluster-1
```

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 and a high confidence and Rule#4 and #Rule5, with a high support and low confidence. Because we are interested not only in having high performance (rules with high confidence), but certainly also in wide-coverage general rules which may provide new useful knowledge, we inspect rules Rule#4 and Rule#5 more closely. Using the same reasoning for the rules induced to capture cluster-3, we focus on Rule #2, Rule#3 and Rule#4.

The wide-coverage rules tell us that if a patient has three or less diagnoses, and does not have diagnosis 585 (renal failure), it is likely that he/she will be in cluster-1: a "moderately complex" patient. In contrast, if a patient has diagnosis 585 (renal failure), it will be a "complex" patient. Also, according to Rule #4 for cluster-3, if the number of diagnosis is higher than 2, it will estimated to be a "complex" patient. If the patient does not have diagnosis 585, 401, 272 and heart problems, has in total three diagnosis and is a woman, she has some chance to be a "complex" patient.

The rules provided by this predictive model do not explain all. As a second illustration, we inspect the predictive model from experiment "all diagnosis" with LOG_VAR_3. A selection of the rules with

support higher than 20 instances and confidence higher than 0.6 is shown in Figure 5.

**Figure 5: Predictive rules from experiment "chronic diseases" with clustering model LOG_VAR_3.**

```
Rules for cluster-1:
Rule #1 for 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, 0.612)
Rule #5 for cluster-1:
 if  d585 == 0 and  d2507 == 0 and  d429 == 0 and  C_sec_diag =< 2 and  d286 == 0 and  d250 == 0 and  d425 ==
 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, 0.638)
Rule #7 for 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, 0.686)
Rule #9 for 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 then -> cluster-1 (56, 0.625)

Rules for cluster-3:
Rule #1 for cluster-3:
  if  d585 == 1 and  d442 == 0 and  d429 == 0 and  d444 == 0 then -> cluster-3 (74, 0.703)
Rule #2 for cluster-3:
 if  d585 == 0 and  d2507 == 1 and  C_sec_diag =< 7 and d414 == 0 and d250 == 1 then -> cluster-3 (53, 0.792)
Rule #8 for 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, 0.787)
Rule #13 for 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, 0.915)
Rule #14 for 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, 0.621)
Default : -> cluster-1
```

Among the rules induced for cluster-1, let's look at Rule#1: if a patient has diagnosis 250 (diabetes) (and does not have the other eight specified diagnosis), has two or less than two diagnosis and age more than 55, it's 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 250 the diagnosis 2507 (diabetic foot), it will be assigned to cluster-3, which is the cluster for "complex" patients. Subsequently, the number of diagnosis will be higher, which is also according to the rule (number of diagnosis C_sec_diag =<7). Thus, our model contain a rule that is able to "send" the patient to the right cluster, when an additional diagnosis become known.
Another meaningful rule is Rule#1 for cluster-3, which says that if a patient has diagnosis 585 (renal failure) and do not have the other three specified diagnosis, he/she will be a "complex" patient. Thus, this rule provide 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 can be clustered in homogeneous logistic 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 logistical point of view and not only from the medical point of view.

The two considered approaches, i.e. clustering on logistical variables, and clustering based on latent factor extracted from logistical variables, both lead to three main clusters, of which two hold clearcut groups of patients: one can be labeled "moderately complex" patients, while the other holds "complex" patients.

The remaining third cluster contain 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 logistical 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 ruleset. Possible tests based on this feature are removed from the rules because they don't 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.

Second, we were 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 health care 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 diagnosis" with clustering model LOG_VAR_3, shown below (Figure 6):

**Figure 6: A selection of predictive rules from experiment "all diagnosis" with clustering model LOG_VAR_3.**

```
Rules for cluster-1:
Rule #1 for 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, 0.612)
....
Rules for cluster-3:
Rule #1 for cluster-3:
  if  d585 == 1 and  d442 == 0 and  d429 == 0 and  d444 == 0 then -> cluster-3 (74, 0.703)
Rule #2 for cluster-3:
 if  d585 == 0 and  d2507 == 1 and  C_sec_diag =< 7 and d414 == 0 and d250 == 1 then -> cluster-3 (53, 0.792)
```

Rule #1 for cluster-1 say that a patient is "moderately complex" if he/she doesn't have diagnosis 585 (renal failure), 2507 (diabetic foot), 429, 286, but has 250 (diabetes) and C_sec_diag =< 2. In contrast, using Rule #2 for cluster-3, a patient is estimated to be "complex" if he/she additionally has diagnosis 2507 (and not diagnosis 585 and 414), increasing the number of diagnosis, and C_sec_diag =< 7. Rule #1 for cluster-3 expresses that as soon as a patient has diagnosis 585 (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 enough detailed 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 clearcut clusters, namely "complex" and "moderately complex" patients. Second, the rules for diagnosing patients to clusters that are automatically induced by data mining techniques 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 to predict better the logistic clusters. First, 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 can be made and patients can be treated in the right way faster. Information of this type can be gathered throughout time, and changes in patient groups and treatments could automatically be discovered and relayed back to the logistic management to inspect whether the new data warrant new changes.

As a second branch of future work, we want to determine automatically for each logistic cluster the most frequent patterns of visits to specialists. These patterns could provide more clues for potential reorganizations of a hospital in the direction of specialty-oriented units.

# References


[1] Bertrand J.W.M., Wortmann J.C., Wijngaard J. Production control. A structural and design oriented approach. Elsevier, Amsterdam, 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. http://www.casemix.org.

[4] Cios K.J., Teresinka A., Konieczna S., Potocka J., Sharma S. Diagnosing Myocardial Perfusion SPECT Bull's-eye Maps – A Knowledge Discovery approach, IEEE Engineering in Medicine and biology Magazine, special issue on Medical Data Mining and Knowledge Discovery, 2000; 19(4): 17-25.

[5] Dilts D., Khamalah J., Plotkin A. Using Clustering Analysis for Medical Resource Decision Making, Medical Decision Making 1995;15 (4): 333-347.

[6] Fetter R.B. The new ICD-9-CM diagnosis-related group classification scheme, HCFA Pub. No. 03167, Health Care Financing Administration, Washington: U.S. Government Printing Office, 1983.

[7] Fetter R.B., Averill A. Ambulatory visit groups: a framework for measuring the productivity in ambulatory care. Health Services Research 1984; 19: 415-437.

[8] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective, Artificial Intelligence in Medicine, 2001; 23: 89-109.

[9] Lavrač N. Selected techniques for data mining in medicine, Artificial Intelligence in Medicine, 1993; 16:3-23.

[10] Quinlan J.R. C4.5: Programs for Machine Learning, Morgan-Kaufmann, 1993.

[11] Miksch S. Plan management in the medical domain, AI Communication, 1999; 12: 209-235.

[12] Ploman M. Choosing a Patient Classification System to describe the Hospital Product, Hospital and Health Services Administration, 1985; May-June:106-117.

[13] Spyropoulos, C.D. AI planning and scheduling in the medical hospital environment, Artificial Intelligence in Medicine, 2000; 20: 101-111.

[14] Vissers J., Patient flow based allocation of hospital resources, Doctoral thesis, Technical University of Eindhoven, the Netherlands,1994.

[15] Vries, G. de, Bertrand J.W.M., Vissers J.M.H. Design requirements for health care production control systems, Production plnning & control, 1999; 10 (6): 559-569.

[16] Vries, G.G. de, Vissers J.M.H., Vries G. de. The use of patient classification systems for production control of hospitals. Casemix Quarterly, 2 (2):65-70.

[17] Vries, G.G. de, Vissers J.M.H., Vries G. de. Logistic control system for medical multi-disciplinary patient flows. In: Monitoring, evaluating, planning health services, ORAHS'98, 1998; 141-151.