Towards decision support for a home care services platform

Jan-Willem van 't Klooster
University of Twente
Faculty of EEMCS
PO Box 217
7500AE Enschede, The Netherlands
j.w.vantklooster@utwente.nl

Catherine Combes
Université de Lyon
UMR CNRS 5516 - HUBERT CURIEN Laboratory
F-42023 Saint-Etienne, France
catherine.combes@univ-st-etienne.fr

Bert-Jan van Beijnum
University of Twente
Faculty of EEMCS
PO Box 217
7500AE Enschede, The Netherlands
b.j.f.vanbeijnum@utwente.nl

ABSTRACT
It is believed that ICT-mediation for home care services increases patient empowerment, independency, self-efficacy and quality of life. Providing elderly people with tailored care services allows us to learn from patient data to predict future care needs. In this article, we demonstrate the contribution of machine learning to homecare services, using data collected by a home care services platform. As an actual case, we show how simulated medication compliance can be measured and modeled using clustering and regression techniques. The approach is validated using data from French nursing home databases. The results show that it is possible to classify situations in elderly healthcare, and schedule resource planning according to expected health problems.

Categories and Subject Descriptors
H2.8 [**Data mining**]; J.3 [**health**]:

General Terms
Algorithms, Design, Simulation.

Keywords
Data mining in healthcare systems, machine learning.

1. INTRODUCTION
The aging population is an omnipresent problem in western countries. For example in the Netherlands, one of 4 to 5 working class people would need to work in healthcare in 2040 to provide the current level of care [12]. As this is undesirable from an economic and social perspective, it is important to innovate care and to put clients more into a self-management position, using ICT to support them in their daily life and only call for professional aid in case of necessity. This is also desired by most elderly as they express to age at as independent as possible.

Clearly, (electronic) care services can contribute to this desire and this problem but in general they are expensive, they are not well integrated and not well tailored to specific needs [9]. Healthcare institutions should systematically acquire the information needed to make decisions and to react quickly. If we focus our interest on nursing homes, we see elderly people pathology development induces health care and social needs, and involves problems of management. These problems include organizing, leading, acquisition and allocation of resources, but also controlling and managing activities. Resource allocation (material, human and financial) and the obligation to react quickly in the relevant cases can however be improved by ICT services.

In the U-Care project [8], these ICT services are subject of study. Hereto, an ICT system is developed to provide tailorable home care services, based on a services composition approach. This way it is possible to realize personalized care services based on standard building blocks inside the system, or using third-party service offerings. The case that is used in this paper is one example of such a composition: an agenda service and a reminder service are orchestrated to support taking medication once a day, the medication diary is monitored accordingly, and an alert is triggered to a caregiver in case of a too bad compliance.

Relatively different from similar projects in this domain is the learning aspect that is the subject of this article: The medication compliance data and agenda activities are used to learn patient behavior over time. We are interested in learning from patient data in the developed system, to enable decision support and planning of resources. This project is realized in close collaboration with the nursing home Parc Hoogveld in Sittard, The Netherlands. The resulting data, stored in the system, is used for patient proofing, planning and forecasting long-term care as demonstrated in the next section. First however, a more general context of elderly care is given to show the typical issues.

Long-term care is needed when people have a chronic illness or disability which requires assistance needs in Activities of Daily Living (ADL) such as help performing for washing, dressing, using the toilet, transferring (to or from bed or chair), caring for incontinence, eating, taking medication and the ability to alert.

To fully explore the opportunities for our approach, we propose a decision support system based on data collection in these areas and data mining approaches, in order to support operational and tactical decisions, and to find relevant patterns among the measured data.

Permission to make digital or hard copies of all or part of this work for personal or classroom m use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WI&C'12, April 16, 2012 Lyon, France
Copyright © 2012 ACM 978-1-4503-1189-2/12/04... $10.00.

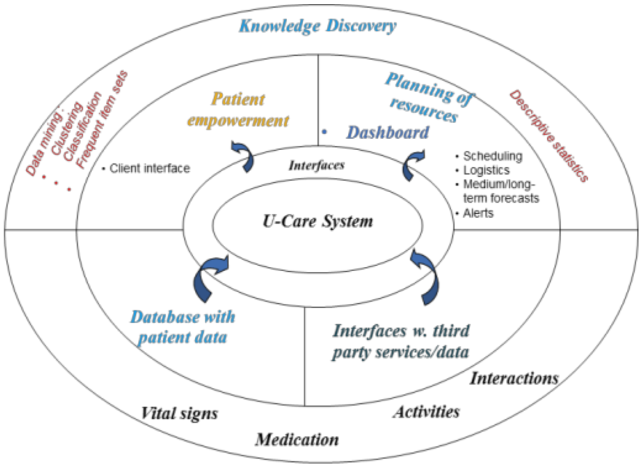


Figure 1. General architecture of the U-Care ecare system

The approach is as follows:

(1) requirements analysis for system components;

(2) development and testing;

(3) simulation of elderly people data;

(4) clustering of patient data;

(5) forecasting and pattern finding;

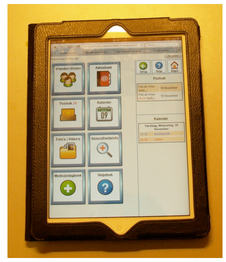
(6) presentation of results in the form of an interactive dashboard for caregivers and management.

Step 1-5 are presented in this paper as follows. First the system is briefly explained in section 2. Section 3 discusses the methods used in the data mining and why these methods are used. Section 4 presents the results of a simulation study. Then a validation of the method using existing data from French nursing homes is performed. These data are the results of staff assessments, rather than measured by an electronic care (e-care) system, but nevertheless show that the same methodology is still suitable because both sources concern the same domain and data are in the same format. Conclusions and an outlook are finally presented in Section 5 and 6.

2. HOME CARE SERVICES SYSTEM

The general architecture of the developed care system is presented below.

The system is fed with information about among others vital signs, medication, activities and interactions from either the database of the patient system (described below) or web services interfaces with third party services. This information is used to inform patient via the client interface (using a tablet computer as shown in Figure 2a), and to provide a dashboard to caregivers (as shown in Figure 2b). The following describes the components in more detail.



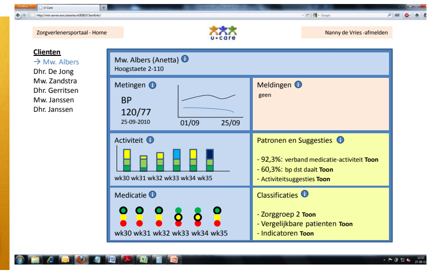


Figure 2a (left): Client access for vital signs, medication, activity and interaction services using touchscreen device.

Figure 2b (right): Caregiver dashboard access for monitoring of vital signs, medication, patterns and classifications (offered as a web application).

2.1 Patient system

The patient system consists of a touch screen device (touch screen pc or tablet) running a set of web-based applications, offered according to the client needs. Example applications are a medication diary, a calendar and a friend finding application based on a community.

The system has a set of adapters to connect to third party services via webservices (WSDL) technology. This way not only built-in service data can be used for knowledge extraction but also the data provided by the third party service. An example is the availability of blood pressure measurement data using the Omron device of the MobiHealth (www.MobiHealth.com) telemonitoring service, which is pushed to the ucare system when a patient made a measurement [15]. Caregivers may configure a reminder to the client to perform this kind of self-measurement.

2.2 Caregiver system

The caregivers utilize dedicated software to set reminders and other workflow processes based on templates, so as to configure the services available to the client. They can thus tailor the system according to client needs, and monitor the performance over time. For example, they can set reminders for medication to be taken based on an intake schedule or add social activities of a certain type (eg bowling) to the caretakers' agenda.

For caregivers it is interesting to see which patients are doing well and which are not, so a dashboard (as shown in Figure 2b) is proposed to provide the information per patient. On a larger scale, it is interesting for planning and logistic purposes to classify patients into groups; to predict the class of new cases; to perform outlier detection; and to find rules that explain patient behavior. These contributions are discussed in the following of this paper.

In conclusion, the system presented in figure 1 is a healthcare information system targeted at providing care services to clients. The personalization and management of these care services is exposed to the caregiver.

3. METHODS

This section describes the methods used in the data mining approach. A simulation is made to fill the system prior to actual patient tests which is due to practical and clinical constraints not possible at this time for the desired number of users. As a result, first the simulation approach is described and consecutively the data mining methods for clustering and regression purposes as corresponding to the formulated goals in section 1.

3.1 Pseudorandom data generation

Let a be a day ∈ { Monday , … , Sunday }. For n=126 patients, we generated 4 weeks of medication compliance data where the compliance *c(i,a)* of patient *i* on day $a \in \{0,1,2\}$, i.e. bad, improvable, or good compliance for that day. Boundaries for these ordinal values may be set in the system by the responsible caregiver per client. It is important to notice that we use ordinal values. Although this certainly limits the possible data mining methods, and we have to take the distance measure into account, we can now work with terms such as good compliance and bad compliance based on what caregivers define and hence they can define it different from client to client.

We introduced random behavior for Monday and Thursday and introduced the notation of 4 types of patient compliances: average patients, good compliant patients, good-minus compliant patients, and weekend-minus compliant patients, as found in the attached datafile 1. For all categories weekend-data were varied between 0, 1 and 2. The weekend days were hence instantiated orthogonal, i.e. for each patient type all combinations of $c_{sat} \times c_{sun} = |c|^2 = 9$ are simulated.

Clustering based on coupling b-coloring of graphs [4,5] and k-medoids [2] was then executed to find patient categories and extract knowledge on resources (eg medication intake supervision) planning. This method consists of b-coloring the most distant nodes in the vector space in order to estimate the number of clusters and then k-mediods is performed based on the number on the dominant colors in this graph. It gives better results than PCA- (Principle Component Analysis) based clustering if the data is more unrelated. In fact, k-mediods is a general version of k-means which works with any distance measure, whereas k-means only works for Euclidian distances (and hence not unimplicitly for ordinal data). The drawbacks of these methods (I: initialization-sensitivity, II: possibility to converge to local optimum: III: desired number of clusters has to be given (supervised method)) can be tackled by performing graph b-coloring first, in order to identify the number of clusters as input to k-mediods.

3.2 Data mining description

For clustering we need to define a measure of distance between instances based on the type and value of their variables. We use the Generalized Measure Distance (GMD) [7] as a distance measure in favor of Euclidian distance as a distance metric. GMD can be used for ratio, interval and ordinal scales. Though we set the weight for all variables to 1, the measure allows for differences in weight which potentially allows different variables related a client to be weighted differently.

Logistic regression

Logistic regression is executed afterwards in order to reveal the rules to match patient into the found categories and the most contributing variables. Logistic regression analysis is performed using Weka [13]. This logit-based method, where [11]

$$\mathit{logit}\ \mathrm{p}(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

yields us the likelihood of a new instance belonging to the found classes and the contribution of each variable x_i to the decision.

We are able to estimate the membership of a cluster C_i regarding the value of the variables $x_1 \ldots x_n$ using the following formula:

$$P\,[Y = C_i / X_1,\ldots,X_n] = 1 / [\ 1 + exp(-(\beta_0 + \Sigma_i \beta_0 X_i)\ ]$$

The experiment was concluded by performing frequent item-set analysis to find the most prominent associations in the simulated data.

We used Tanagra to find association rules using the apriori algorithm [Liu 2007]. As apriori algorithm only works with Boolean values for each variable, we have to recodify the 3-valued variable c into 3 Boolean valued variables c’, e.g.:

$$\{ c_x = 2 \} \equiv \{ c'_{x,0} = 0 \ ; c'_{x,1} = 0 \ ; c'_{x,2} = 1 \}$$

where *x* is a day *a*.

4. Results

We want to find the actual clusters hidden in the medication data. Two methods are tried; PCA on linear correlation between variables, and b-coloring of graphs coupled to k-mediods algorithm. After that, the regression results and patterns found using frequent itemset analysis are shortly discussed.

4.1 Clustering with PCA.

This approach has been proposed in [1]. The proposed algorithm is based on PCA and partitions.

The principle of PCA is to compute the matrix Õ which corresponds to the centered and reduced data from the matrix O (obtained by subtracting the corresponding mean and by dividing by the corresponding standard deviation of each random variable). Afterwards, we compute the correlation matrix C from Õ in order to find the eigenvectors and the eigenvalues. From the first two eigenvectors we verify if there exists a linear correlation i.e. the characteristics of the two first component diagram is that we observe parallel lines (visual analysis concerning the first two principal components) implying that the first two principal components have a high cumulative correlation in order to retain the maximum amount of information. After, we execute the partition algorithm based on the first two eigenvectors. We compute the P_i coordinates of the first two components (p_{i1} , p_{i2}) and we apply rotation and projection qi on the corresponding axis. Afterwards, we compute d_k, the Euclidian distance from each qk to its successor depending on the sorting and each bisector of the line segment corresponding to nc-1 longest distances is the cutting between two clusters.

If these preconditions are not satisfied, it is not possible to use this algorithm because we have to take into account more than the first two components. In this case, we have the same problem as with the k-means approach. We have to correctly identify the kernels of each cluster, and PCA clustering becomes similar to k-means clustering [3]. We try to explore it by using connected components or the k-medoids algorithm (a clustering algorithm related to the k-means algorithm and the medoidshift algorithm, dedicated to ordinal data).

In the studied case, we first we conducted a cluster visualization on all the days in the week, seeing that the first two component explain 52,6% of correlation. So there is not a good linear correlation and PCA-based clustering doesn’t yield us valuable results. In fact, it is not surprising as 2 days are filled random, and the 2 weekend-days are orthogonal for all instances. Figure A.1 shows the result.

4.2 Coupling b-coloring of graph and k-means on all variables

As we found that the data is relatively uncorrelated concerning all the variables, PCA clustering is not appropriate. Hence we use a new approach based on coupling b-coloring of graph and k-medoids algorithm. This approach has been proposed in [2,4,5].

The principle is the following. Graph coloring is a special case of graph labelling. The proper b-coloring problem [6] is the assignment of colors to the vertices of one graph with two conditions:

(1) Adjacent vertices have different colors;

(2) For each color, there exists at least one vertex having its neighboring vertices in each other colors. This vertex is called a dominating vertex.

The advantage of the b-coloring is that we automatically find the number of clusters which corresponds to the number of different colors of the dominating vertices. We can also identify the "best representative object" of each cluster regarding the vertices having the same color. That way, we optimize the choice of these medoids with respect to k-medoids choosing datapoints as centers. Moreover we obtain the number of clusters (this input number is necessary for the k-medoids algorithm) instead of trying manually. We experiment the approach on the simulated data concerning medication. Table 1 shows the results of the clustering and a short description of each of the clusters found.

Table 1. Clustering description of 126*4=504 patient weeks, n=cluster size;%=percentage of total; d=dispersion;s=silhouette

Cluster # (n;%;d;s)	{ Mo , … , Su }	Description
0 (94;18,7%;.08;.7)	{1,1,1,1,1,1,1}	Average patient
1 (76;15,1%;.07;.6)	{2,2,2,2,2,2,2}	Compliant patient
2 (83;16,5%;.08;.4)	{2,2,2,2,2,2,1}	Compliant patient / Sunday
3 (107;21,2%;.1;.2)	{1,2,2,2,2,2,2}	Compliant patient / Monday
4 (34;6,8%;.1;.4)	{2,2,2,1,2,0,1}	Weekend incompliant
5 (66;13,1%;.1;.4)	{2,2,2,0,2,2,2}	Thursday incompliant
6 (44;8,7%;.1;.3)	{0,2,2,2,2,1,1}	Sat-Mon incompliant

4.3 Logistic Regression

We used Weka [13] to search a classifier for all variables. The classifier is listed below. A maximum of 5 days is needed to predict the class so based on the training set it is possible to anticipate resources on expected compliance issues likely to occur later in time (99.4% accuracy). The confusion matrix belong to the found regression model is listed in Appendix A.

Listing 1. Regression model with Variables and Classes as in Table 1.

```
Class 0 :
37.64 + [MON] * -1.8 + [TUE] * -1.32 + [WED]
* -8.64 + [FRI] * -8.22 + [SAT] * -6.36
Class 1 :
-35.15 + [MON] * 5.9  + [THU] * 6.47 + [SAT]
* 0.57 + [SUN] * 6.77
Class 2 :
-8.77 + [MON] * 4.66 + [THU] * 2.32 + [SAT]
* 0.91 + [SUN] * -3.61
Class 3 :
-4.47 + [MON] * 3.79 + [THU] * -7.99 + [SUN]
* 3.56
Class 4 :
-6.73 + [MON] * -4.17 + [TUE] * 0.62 + [THU]
* 2.39 + [SAT] * 2.74 + [SUN] * 1.82
Class 5 :
7.94 + [MON] * 3.93 + [THU] * -2.44 + [SAT]
* -9.39 + [SUN] * -2.03
Class 6 :
16.07 + [MON] * -7.23 + [TUE] * -0.96 +
[THU] * -1.04 + [FRI] * 1.16 + [SAT] * -6.96
```

4.4 Pattern finding

Table 2 gives somes examples of patterns found using Apriori within Tanagra. Using this algorithm, it is possible to find patterns among measured data and the significance of these patterns.

An example of frequent itemset analysis, showing the pattern found between intake compliance on Tuesday, Wednesday and Friday. In the dashboard case, this algorithm is used to show the most eminent patterns found.

Table 2. Some Pattern examples using Apriori.

```
Antecedent  Consequent   Support      Lift    Confidence
FRI=true  <- WED=true  (0.9523810, 0.9756098, 0.9761905)
WED=true  <- FRI=true  (0.9523810, 0.9756098, 0.9761905)
WED=true  <- TUE=true  (0.9523810, 0.9756098, 0.9761905)
TUE=true  <- WED=true  (0.9523810, 0.9756098, 0.9761905)
FRI=true  <- TUE=true  (0.9523810, 0.9756098, 0.9761905)
TUE=true  <- FRI=true  (0.9523810, 0.9756098, 0.9761905)
…
```

4.5 Discussion

From the simulated data of medication, we can find some interesting properties. Clearly there are some patients that need supervision every day: class 0 represents 20% of the population. Cluster 1, 15% of the simulated population doesn't need attention. Cluster 3 (21%) should be checked Mondays. Cluster 4 (6,8%) needs supervision in the weekend and cluster 5 (13%) needs supervision on Thursday. Using the regression model built, we see that is possible to anticipate of compliance problems that are, based on the training data, likely to occur later in time (as it's likely to know the class after a maximum of 5 days). Of course in practice, new data should be added continuously to reinforce the classifier and the regression model.

As Weka (Java) can be integrated in a J2EE service oriented architecture (SOA) environment, it is possible to chain the presented data mining actions such that a dashboard (like presented in section 2.2) can display the results to the end users.

5. Validation

We tried to validate the experiment in different ways.

First we compared the b-coloring / k-mediods based clustering with PCA-based clustering, this yielded the same number of clusters (7), but worse cluster quality (53% explained in 2 variables) as discussed in section 4.

We compared our simulated data also with data recorded from real patients using A.G.G.I.R., ("Autonomy-Gerontology-Group-Iso-Resources") a French codification system to assess patients in nursing homes. We performed the same cluster analysis using this data (discussed in the next sections). Additionally we tried to compare medication data, which was not possible as 99% patients in the A.G.G.I.R. database suffered from medication problem, so there was no distribution among this variable. This is explained in the following part of the report.

5.1 Comparison study

Long-term care is needed in patient suffering from chronic illness or disability. These conditions lead to a need for assistance with Activities of Daily Living, such as help for washing, dressing, using the toilet, eating, transferring, etc. Disability is originally defined as any restriction or lack of ability (resulting from an impairment) to perform an activity in the manner or within the range considered normal for a human being.

The term "disability" is now used as an umbrella term covering any or all of the following components: impairment, activity limitation and participation restriction, as influenced by environmental factors [14].

Dependence evaluation in France is carried out using a specific scale called the A.G.G.I.R scale. Figure 1 presents the A.G.G.I.R scale. It is a scale which is based on a set of items in order to determine the autonomy-disability for each individual in Activities of Daily Living. In nursing homes, only the first eight items are observed:

(1) Coherence,

(2) Orientation,

(3) Washing,

(4) Dressing,

(5) Food,

(6) Toilet,

(7) Transferring,

(8) Moving.

The evaluations are made by the resident doctor in collaboration with the medical staff.

An item can be evaluated using one or more of the four adverbs:

- Spontaneously,
- Usually,
- Entirely,
- Correctly.

The codification is the following (see figure 3):

- If all adverbs are applicable, the code is A;
- If < 4 adverbs are applicable, the code is B;
- If 0 adverbs are applicable, the code is C.

The aim is to find feature-patterns related to the autonomy-disability level of elderly people from the data clustering: a common technique for data analysis.

5.2 Experiment

The aim of the study is, again, to identify differences between feature-patterns related to the autonomy-disability level of elderly people in the different clusters obtained from residents suffering from dementia syndromes (1,010 observations) related to cognitive problems and to apraxia. The same method (b-coloration of graphs and k-medoids) as described in 4.2 is used on these data, which is also codified in 0, 1, 2 (again, resp. A, B, C).

Clustering is performed on residents living in nursing homes. The results are shown in Appendix B, revealing that this method succeeds in clustering in real life cases [1].

A.G.G.I.R.

For each item, tick the box(es) corresponding to the number of the adverb(s) which best describes the person's ability:

1. Spontaneously
2. Usually
3. Entirely
4. Correctly

Mental and corporal autonomy

		1	2	3	4	Code
5.1 Coherence		☐	☐	☐	☐	☐
5.2 Orientation		☐	☐	☐	☐	☐
5.3 Washing	upper body	☐	☐	☐	☐	☐
	lower body	☐	☐	☐	☐	☐
5.4 Dressing	upper body	☐	☐	☐	☐	☐
	middle body	☐	☐	☐	☐	☐
	lower body	☐	☐	☐	☐	☐
5.5 Food	to serve	☐	☐	☐	☐	☐
	to eat	☐	☐	☐	☐	☐
5.6 Toilet	urinary	☐	☐	☐	☐	☐
	faecal	☐	☐	☐	☐	☐
5.7 Transferring (to or from bed or chair)		☐	☐	☐	☐	☐
5.8 Moving indoors		☐	☐	☐	☐	☐
5.9 Moving outdoors		☐	☐	☐	☐	☐
5.10 Ability to alert		☐	☐	☐	☐	☐

Social and domestic autonomy

	1	2	3	4	Code
6.1 Management	☐	☐	☐	☐	☐
6.2 Cooking	☐	☐	☐	☐	☐
6.3 Cleaning	☐	☐	☐	☐	☐
6.4 Transport	☐	☐	☐	☐	☐
6.5 Buying	☐	☐	☐	☐	☐
6.6 Medication	☐	☐	☐	☐	☐
6.7 Free time activities	☐	☐	☐	☐	☐

The corresponding code is :
A: spontaneously, usually, entirely and correctly (the whole set of adverbs are checked).
B: some adverbs are checked.
C: no adverb are checked.

Iso-resource group ☐

Figure 3. A.G.G.I.R. Nursing home client assessment.

6. CONCLUSIONS

We showed that different data mining techniques can be used beneficially in the analysis, classification, prediction and suggestion of patient information using data mining on telecare system data.

The system that acquires this information has been developed and will be validated in due course. It was discussed in section 2.

We simulated 126 patients using this system regarding their medication compliance. On this simulation, we performed clustering, model-making and pattern finding. The clustering clearly indicates patient classes with patients that need extra supervision on certain days. A regression model was constructed to define where new instances belong to. Finally, we demonstrated that patterns can be found in this data using frequent item set mining.

The novel clustering method that was utilized in the simulation was validated with AGGIR data from a French nursing home. Given the presented ecare system and the data mining methods that have been presented, it becomes feasible to present the clustering, prediction and pattern finding results in an intelligent web environment to support caregivers and care management in their decision making.

7. FUTURE WORK

Three items are interesting in the short and medium future. Firstly, real data should be gathered using the ecare system to validate this simulation, but also to test this approach beyond solely medication compliance measures. This will be done in the course of 2012, when a pilot is organized in the nursing home Parc Hoogveld, Sittard, to validate the ecare system among clients and caregivers. The data from that pilot will be used in this data mining approach, which implies that the clients and caregivers have to be aware of the fact that their use of the system is logged and used for individual monitoring, but also, in anonymed form, for management issues such as the clustering and model-making discussed in this paper. To this end, it has been deciced that the infrastructure of the ecare system will be located behind the firewall of the nursing home, maximing the security of the data gathering.

Secondly, this data should be used for longitudinal monitoring. For this, the proposed dashboard needs to be tested and it should include the regression model and best suggestions based on a fine-tuned Frequent Item-Set Mining to show suggestions to the caregiver based on the client performance. The Apriori algorithm is useful to find the frequent item sets on these measured variables in the system. It is interesting to retrieve frequently occurring relations among the different health parameters. A selection is then to be displayed in the proposed dashboard.

Thirdly, on a longer timescale, we can theoretically predict if patients will move to a different class using a Markov Modelling approach. This should result in suggestions for caregivers on what to do in these cases.

8. ACKNOWLEDGMENTS

This work was sponsored by the IOP GenCom U-Care project (http://ucare.ewi.utwente.nl), sponsored by the Dutch Ministry of Economic Affairs under contract IGC0816, and financial support of the Région Rhône-Alps, France.

9. REFERENCES

[1] Combes, C. and Azema, J. 2012A. Clustering using Principal Component Analysis Applied to Autonomy Disability of Elderly People, to appear in *Decisional Support system, Special issue Healthcare*.

[2] Combes, C. and Azema, J. 2012B. Clustering in coupling b-coloring of graph and k-medoids: an application in addictilogy. *Submitted to MOSIM 2012*, June 6-8, Bordeaux.

[3] C. Ding C., X. He, K-means Clustering via Principal Component Analysis. *Proc. of Int'l Conf. Machine Learning (ICML 2004)* 225-232. http://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf.

[4] H. Elghazel, V. Deslandres, M.S. Hacid, A. Dussauchoy and H. Kheddouci. A New Clustering Approach for Symbolic Data and its Validation: Application to the Healthcare Data, *In the proceedings of 16th International Symposium on Methodologies for Intelligent Systems (ISMIS 2006)*, Bari, Italy, pages 473-482, 2006.

[5] H. Elghazel, K. Benabdeslem and A. Dussauchoy. Constrained Graph b-coloring based Clustering, *In the proceedings of 9th International Conference on Data Warehousing and Knowledge Discovery (DAWAK 2007)*, LNCS Nº4654 - Springer Verlag - ISBN: 978-3-540-74552-5, Regensburg, Germany , pages 262-271, 2007.

[6] W. Irving, D. F. Manlove, (1999). The b-chromatic number of a graph. *Discrete Applied Mathematics,* Vol. 91, 127-141.

[7] K. Jajuaga, M. Walesiak, A.Bak. On the general distance measure. *In Exploratory data analysis in empirical research, Gesellschaft für Klassifikation*. Jahrestagung, Manfred Schwaiger, Otto Opitz (ed.). Springer, 2003. http://books.google.nl/books?hl=nl&lr=&id=Y_pWuH_CMfQC&oi=fnd&pg=PA104&dq=walesiak&ots=HsXLDK4v5x&sig=llDRKrSNTJE0QTw6Lnhp3X64TBE#v=onepage&q=walesiak&f=false

[8] van 't Klooster, J.W.J.R. and van Beijnum, B.J.F. and Pawar, P. and Sikkel, K. and Meertens, L.O. and Hermens, H.J. (2011) Virtual Communities For Elderly Healthcare: User-Based Requirements Elicitation. *International Journal of Networked & Virtual Organizations*, 9 (3). pp. 214-232. ISSN 1470-9503.

[9] G. Lafortune , G. Balestat. 2007. Trends in Severe Disability Among Elderly People: Assessing the Evidence in 12 OECD Countries and the Future Implications. *Report of Directorate For Employment, Labour And Social Affairs Health Committee*, DELSA/HEA/WD/HWP.

[10] Liu, B. 2007. Web Data Mining. *Exploring Hyperlinks, Contents, and Usage Data*. Ch 1-4. Springer-Verlag.

[11] L. Rouviere. *Regression sur variables Categorielles*. Universite Rennes 2, 2008.

[12] Schippers, E.I., Veldhuijzen van Zandten, M.L.L.E. (2011). *Arbeidsmarktbrief, Brief aan de Tweede Kamer der Staten Generaal.* 4 Maart 2011.

[13] WEKA, http://www.cs.waikato.ac.nz/ml/weka/. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1.

[14] WHO, International Classification of Functioning, Disability and Health. Geneva. 2001.

[15] Zarghami, A., Zarifi Eslami, M., Sapkota, B., van Sinderen, M. Toward Dynamic Service Provisioning in the Homecare Domain. *International Workshop on Designing and Integrating Independent Living Technology, DIILT'11*, Dublin, Ireland, May 2011.

Appendices

A. Logistic regression.

Confusion Matrix.

```
  0   1   2   3   4   5   6  < classified as

 96   0   0   0   0   0   0 |   0 = class 0
  0  76   0   0   0   0   0 |   1 = class 1
  0   0  85   0   0   0   0 |   2 = class 2
  0   0   0  66   0   0   0 |   3 = class 3
  0   0   0   0 119   0   0 |   4 = class 4
  0   0   0   1   0  33   0 |   5 = class 5
  0   0   0   1   1   0  26 |   6 = class 6
```

B. A.G.G.I.R. Clustering

Table B1. Clustering description of 1010 observations regarding dementia patients' coherence and orientation scores. n=cluster size;%=percentage of total; d=dispersion;s=silhouette

Cluster # (n;%;d;s)	Coherence, Orientation
0 (219;21,7%;.002;.98)	{1,1}
1 (90;8,9%;0;1)	{2,1}
2 (536;53,1%;0;1)	{2,2}
3 (43;4,3%;0;1)	{1,2}
4 (122;12,1%;.025;.64)	{1,0}

Table 3. Clustering description of 1010 observations regarding dementia patients' apraxia scores. n=cluster size;%=percentage of total; d=dispersion;s=silhouette

Cluster # (n;%;d;s)	{toilet, habillage, alimentation, elimination, transfer, moving}
0 (65;6,4%;.036;.52)	{2,2,1,2,1,1}
1 (159;15,7%;.028;.56)	{2,2,1,2,2,2}
2 (192;19%;.009;.85)	{2,2,2,2,2,2,2}
3 (56;5,5%;.045;.44)	{2,2,1,2,0,1}
4 (50;5%;.038;.53)	{2,2,1,2,2,1}
5 (83;8,2%;.021;.57)	{1,1,0,0,0,0}
6 (50;5,0%;.092;.054)	{2,1,1,0,0,1}
7 (32;3,2%;.015;.78)	{2,2,2,2,1,1}
8 (52;5,2%;.048;.20)	{1,1,0,1,0,0}
9 (132;13,1%;.028;.39)	{1,0,0,0,0,0}
10 (33;3,3%;.035;.60)	{2,2,1,2,0,0}
11 (62;6,1%;.075;.062)	{1,1,1,2,0,1}
12 (44;4,4%;.096;.13)	{1,1,0,1,1,1}