Bayesian Network-Based Model for the Diagnosis of Deterioration of Semantic Content Compatible with Alzheimer's Disease

José María Guerrero Triviño, Rafael Martínez-Tomás
and Herminia Peraita Adrados

josemaria.guerrero@cpiia.org

Abstract. Alzheimer's Disease (AD) has become a serious public health problem that affects both the patient and his family and social environment, not to mention the high economic cost for families and public administrations. The early detection of AD has become one of the principal focuses of research, and its diagnosis is fundamental when the disease is incipient or even prodromic, because it is at these stages when treatments are more effective. There are numerous research studies to characterise the disease in these stages, and we have used the specific research carried out by Drs. Herminia Peraita and Lina Grasso. The application of Artificial Intelligence techniques, such as Bayesian Networks and Influence Diagrams, may provide a very valuable contribution both to the very research and the application of results. This article justifies using Bayesian Networks and Influence Diagrams to solve this type of problems and because of their great contribution to this application field. The modelling techniques used for constructing the Bayesian Network are mentioned in this article, and a mechanism for automatic learning of the model parameters is established.

Keywords: Bayesian Network, Influence Diagram, Corpus of Oral Definitions, Naive Bayes, Alzheimer's Disease, Cognitive Deterioration.

1 Introduction

As in other fields in the real world, medical diagnosis is not always 100% accurate. In the specific case of Dementia and especially Alzheimer's Disease (AD), its diagnosis is sometimes an extremely difficult task, especially when it is incipient and intensity is only slight[15]. These are the main reasons for using Soft Computing techniques (techniques that enable us to work with incomplete, inexact and uncertain information) to solve this type of problems. Bayesian Networks, in particular, provide a probabilistic model that makes it possible to define the causal relations between the variables explicitly. This causality is assigned a relation force, which is logically determined by the degree of correlation or causality between the variables. Bayesian Networks are extremely useful in response to new cases,and there are Automatic-Learning techniques for both qualitative and quantitative models. These Automatic Learning techniques can enable us to

J.M. Ferrández et al. (Eds.): IWINAC 2011, Part I, LNCS 6686, pp. 419–430, 2011.
© Springer-Verlag Berlin Heidelberg 2011

discover new relations between the variables or new conditional probabilities, as new cases appear. [10,2]

It has been demonstrated that there is a semantic deterioration in the ability to differentiate between living creatures (biological entities) and non-living creatures (non-biological entities) in people suffering certain neurodegenerative pathologies (Alzheimer, Semantic Dementia, Dementia with Lewy bodies, etc.), traumatic pathologies (cranial traumatism), and infectious pathologies (herpes encephalitis). Semantic categories are derived from classifications that are carried out in the world around us and that treat essentially different objects as the same. Thanks to the fact that our semantic memory is organised according to these categories, we can perform a series of important cognitive functions, such as inferring, establishing relations between examples, attributing properties to objects that we do not know, reasoning, all of which is based on a cognitive economy principle. People who suffer specific category deficits execute tasks affecting totally or partially the category domain knowledge of living creatures worse, whereas the object or artefact domain –non-living creatures- is totally or almost totally conserved. There are also a small number of cases where the pattern is the reverse; there is more deterioration in the object or artefact domain, whereas the living creature domain is largely preserved [13].

Thus, the Bayesian Network model constructed in this article aims to diagnose whether the patient suffers cognitive deterioration compatible with AD. This diagnosis is done from a corpus of oral definitions as a methodological tool that has been shown to be very useful to study pathologies in relation to the semantic deterioration of this disease [12]. It is a causal model based on literal definitions of certain semantic categories –of the basic level of categorisation—both of living creatures (dog, pine and apple), and non-living creatures (chair, car, trousers). Patients do some tests where they have to define basic objects. When a patient suffers AD, he suffers from serious cognitive deterioration. The attributes, features or characteristics generated by each patient's definitions are analysed. The underlying logic for analysing the suggested features is in accordance with a model described some time ago in Peraita, Elosúa and Linares (1992) and in keeping with other current works ([3,8,9]).

In the causal model we represent that AD causes a conceptual-semantic-lexical deficit and therefore the Bayesian Network will be able to infer the probability of suffering AD from the degree of conceptual-semantic-lexical deficit. This inference or abductive reasoning starts from some symptoms and searches for the causes that best explain the symptoms. In other words, we start from conceptual-semantic-lexical deterioration and search for the probability that this deterioration is the explanation for suffering AD. Some risk and protection factors, such as educational level, age and sex, will also be taken into account in the Bayesian Network.

Numerous works are currently being done in the field of Bayesian Networks like Early Diagnosis of Alzheimer's Disease. Works on Explanation in Bayesian Networks by [7][1] and other works of interest on Dynamic Bayesian Networks and Learning in Dynamic Bayesian Networks [5] should be highlighted.

Table 1. Abstract of Database of Instances (Linguistic Corpus)

PK	category	taxonomic	types	parts	functional	evaluative	place	conduct	cause	procedural	life cycle	other
1	car	0	1	5	5	0	0	0	0	0	0	3
1	apple	1	2	0	6	0	0	0	1	0	0	0
1	trousers	0	8	0	2	0	0	0	0	0	0	0
1	dog	0	6	0	1	4	0	0	0	0	0	1
1	pine	0	0	0	5	0	0	0	0	0	0	0
1	chair	0	4	0	2	0	3	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...

This work starts from a set of data obtained from the works by [12][14], where semantic category definition features of the linguistic Corpus of healthy subjects and subjects with Alzheimer's Disease are analysed. In other words, we start with a base of instances (table 1 shows an abstract) where the patients' definitions are classified into 11 basic conceptual blocks: taxonomical, types, parts, functional, evaluative, place/habitat, behaviour, causes/generates, procedural, life cycle and others. In other words, each patient's definition is analysed and the number of attributes that he produces is classified into these 11 basic conceptual blocks, with two differential semantic categories, living creatures and artefacts. The following table shows an abstract of the linguistic corpus (database of instances) that this research project has started from.

The research work is in its initial stage. Therefore, this article only describes the model for the diagnosis and the decisions that have been taken for its design. We think that in the very near future we can obtain promising results, especially when we have a wider database of instances, which will enable us to do sufficiently reliable and assessable experiments.

2 Justification for the Technique

Bayesian Networks and influence diagrams offer a number of advantages that make them attractive for use in this application field. The advantages can be summarised in [11,6]:

- Bayesian Networks are Soft Computing techniques and they have to be used because there is currently no deterministic method for diagnosing Alzheimer.
- Bayesian Networks are based on compact graphs and are intuitive of a causal relation between entities of a problem in a specific domain. Other techniques such as Neuronal Networks or Bioinspired Techniques are based on graph representations that are difficult for experts in the field to read.
- Inference is based on the theory of calculating probabilities and the decision theory. It therefore provides a coherent mathematical method to derive conclusions in uncertainty, where multiple sources of information are involved in complex interaction patterns. Other soft computing techniques are also based on complex mathematical models, where the inference explanation is very complex.

- Influence diagrams make it possible to take decisions in a normative way to establish the most appropriate action policy: complementary explorations, pharmacological and cognitive treatments, even for cases where it is not so obvious and the doctor's clinical judgement is unable to find the best solution,
- Decision analysis can explicitly and systematically combine different experts' opinions and experimental data, such as data from studies published in medical literature. [4]
- There are a number of analyses applicable to Bayesian Networks and Influence Diagrams, which provide added value to the technique, like for example, Evidence Conflict Analysis, Sensitivity Analysis (Evidence and Parameters), Value Analysis of the source of information.

There are Bayesian Network Frameworks that perform all or some of these analyses and they can be extremely useful for the problem that we are addressing, because they can also continually validate the model and facilitate the discovery of new relations, analyse new risk factors, new treatments, etc.

3 Proposal for the Diagnosis

As indicated earlier, the Bayesian Network consists of quantitative and qualitative models. For problem modelling four types of variables are used:

- Context information variables or risk factors. It is the information that is present before the problem occurs and that has a causal effect on the problem. In this group of variables we have: age, sex and educational level.
- Information variables representing the symptoms. Variables representing whether the patient has a conceptual-semantic-lexical deficit. These variables analyse the features or attributes contained in a number of semantic category definitions of living or non-living creatures, and other specific ones for each of these categories. The common ones are: taxonomical, functional, part-all, evaluative, place/habitat, types, and the uncommon ones: procedure, behavioural activity, cause/generation and origin. Attribute taxonomy, which acts as a schema or theoretical and methodological evaluation framework for this test, the same as in the second test, can be seen in detail in Peraita, Elosúa and Linares (1992). These variables represent attribute production in each semantic category of each of the objects used (apple, dog, pine, car, chair and trousers).
- Intermediate variables. They are variables that cannot be directly observed whose a posteriori probabilities are not of immediate interest, but they play an important role in achieving correct dependence and conditional independence of the properties and therefore efficient inference. Intermediate variables represent the semantic categories of living creatures and non-living creatures, the semantic content deficit of the different categories, etc. In this type of variables we have: Cognitive Deterioration Living and Non-Living Creatures on the one hand, and Cognitive Deterioration Apple, Dog, Pine, Chair and Trousers, on the other.

– Variables of interest or assumption. We calculate their a posteriori probability from the findings. These variables are: Suffers Cognitive Deterioration (Dementia) and Suffers Alzheimer's Disease.

Discrete Bayesian Network Modelling

Alzheimer and Cognitive deterioration variables of interest have some risk factors represented by the variables Educational Level, Age and Sex. It is worth highlighting that they are risk factors and not causes of the disease, for that reason canonical models (OR/MAX) cannot be used. Furthermore, there is a causal link between "Alzheimer" and "Cognitive Deterioration". According to scientific literature and epidemiological studies, the most common cause of dementia in the European Union is Alzheimer (around 50-70% of cases), other causes of dementia are: multiple cardiac arrest (around 30% of cases), Pick's disease, Lewy bodies and others.

Intermediate variables representing conceptual-semantic-lexical deterioration in the category domain are automatically treated, analysing and interpreting the patient's definitions, distributed in the 11 basic conceptual blocks considered as conceptual components underlying every organisation and representation of object categories (taxonomical, types, parts, functional, evaluative, place/habitat, behaviour, causes/generates, procedural, life cycle and others). Each of these blocks has an identifying lexical label, as indicated in the work by Dr. Herminia Peraita [12,14].

If a patient suffers AD in prodromic or incipient stage, there is a differential deterioration between the semantic categories Living Creatures and Non-Living Creatures. According to Dr. Herminia Peraita's study[12,14], Alzheimer's disease produces cognitive deterioration in Living Creatures before Non-Living Creatures or artefacts. Since AD patients usually have greater damage in the temporal limb areas in the early stages of the disease, they could show selective deterioration for living creatures. As the disease progresses, the damage becomes so omnipresent that mistakes occur in both domains with the same frequency. Therefore, there is a causal relation that qualitative and quantitative models can take into account in this study.

Another important factor that is modelled in the Bayesian Network is the bidirectional correlations between intermediate variables and variables of interest. For example, when a patient produces few attributes in the semantic category of Living Creatures, the probability increases of producing few attributes in the semantic category of Non-Living Creatures when the disease is advanced. Similarly, when the disease is incipient and is Dementia caused by AD, the production of few attributes in the semantic category of Living Creatures increases the probability of producing a greater number of attributes in the semantic category of Non-Living Creatures. In other words, there is a negative causal relation between the two semantic categories when the disease is incipient and a positive correlation when the disease is advanced. We should remember that Bayesian networks do not include cycles. With the intermediate variables it will therefore be possible to represent these circumstances. In fig.1 and fig.2, the modelling techniques used to eliminate the cycles of de bayesian red can be observed.

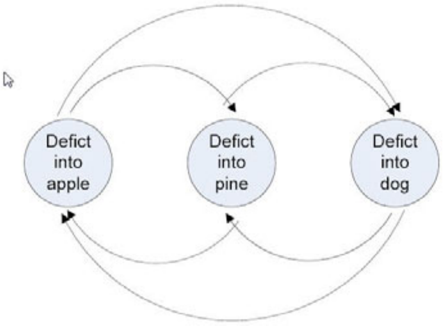


Fig. 1. Conditional dependencies between semantic categories. Graph with cycles.

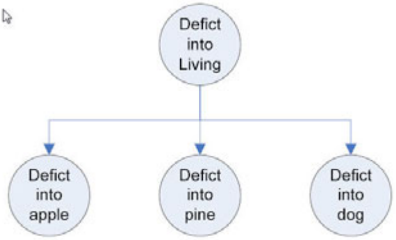


Fig. 2. Conditional dependencies between semantic categories. Graph without cycles.

The fig.3 represents the qualitative modelling of the Bayesian Network. The Construction of the qualitative model is one of the most complex tasks in the modelling of the Bayesian Network. In order to be able to generate the quantitative model automatically, it is necessary to have a large number of representative cases of the population. In many instances it is necessary to have epidemiological studies, and this task is not trivial at all. There are epidemiological studies that provide the distributions of combined probabilities thus: P(DEMENTIA, SEX), P(DEMENTIA, AGE), P(DEMENTIA, EDUCATIONAL LEVEL), P(ALZHEIMER, SEX), P(ALZHEIMER, AGE), P(ALZHEIMER, EDUCATIONAL LEVEL). In this instance it is very useful to use the Naive Bayes simplifier to calculate the conditional probability tables. The Naive Bayesian Method or Naive Bayes starts from the assumption that the diagnosis is exclusive (there cannot be two diagnoses at the same time) and exhaustive (there are no other diagnoses possible). In our model two conditions are fulfilled because we only aim to diagnose cognitive deterioration compatible with Alzheimer at three levels or degrees. With these premises we can use the Naive Bayes simplifier and it is therefore possible to calculate all the conditional probabilities that we need for the model, as can be shown below:

$$P(Alzheimer \mid sex,\ age,\ educational\ level) = a * P(Alzheimer) * P(sex \mid Alzheimer) * P(age \mid Alzheimer) * P(Educational\ Level \mid Alzheimer)$$

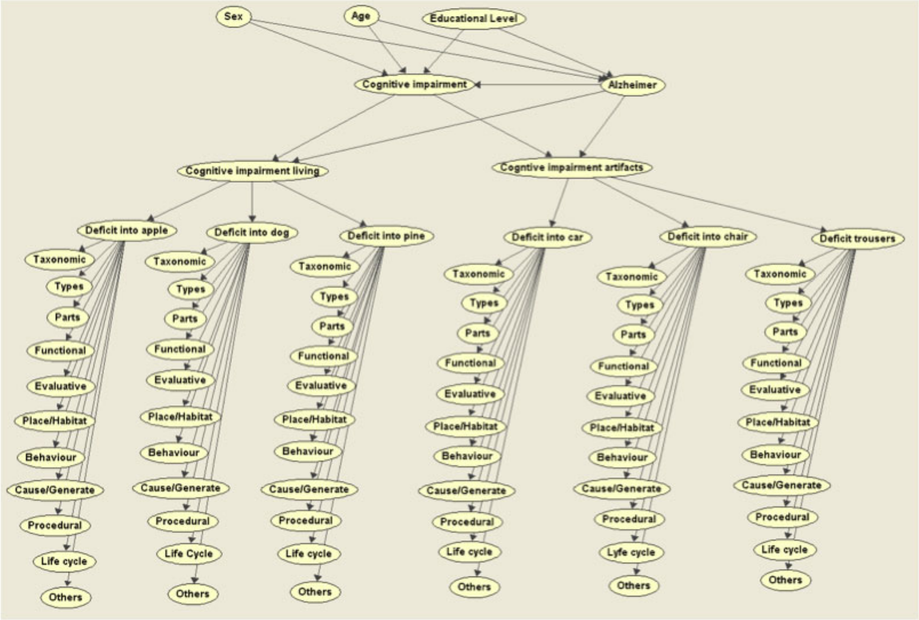


Fig. 3. Discrete Bayesian Network Modelling for the Diagnosis of Deterioration of Semantic Content

Where:

- α is a normalising constant.
- $P\left(sex \mid Alzheimer\right) = \frac{P(sex,Alzheimer)}{P(Alzheimer)}$
- $P\left(age \mid Alzheimer\right) = \frac{P(age,Alzheimer)}{P(Alzheimer)}$
- $P\left(EducationalLevel \mid Alzheimer\right) = \frac{P(EducationalLevel,,Alzheimer)}{P(Alzheimer)}$

For Cognitive Deterioration the same method can be used. Furthermore, an automatic learning algorithm can also be used for the intermediate variables and variables of interest. The total attribute production for each semantic category of each of the objects only has to be added. Then the K-Means algorithm is applied (using WEKA, fig) to create a specific number of data clusters, using the Euclidean distance for this. The K-Means algorithm is based on the patient producing attributes to determine the centroids of each cluster. Each cluster represents each of the states of the variable. Once the centroids of the different clusters have been defined, i.e. the different levels of cognitive deterioration, the database is analysed to calculate the conditional probabilities using the following formulation:

$$P\left(DCx_1 \mid DCSVx_2, FAx_2\right) = \frac{N(DCx_1,DCSVx_2,EAx_3)}{N(DCx_2,EAx_3)}$$

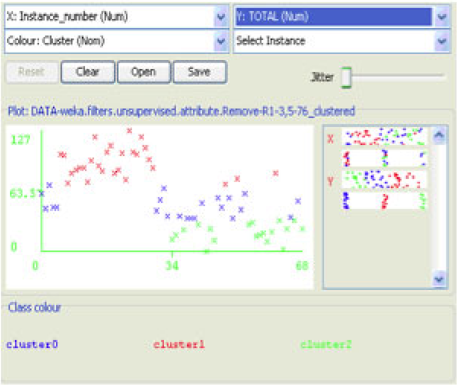
where

$CD \Rightarrow CognitiveDeterioration \Rightarrow \forall x_1 \in \{absent, slight, moderate, serious\}$

$CDLC \Rightarrow CognitiveDeteriorationLivingCreatures \Rightarrow \forall x_2 \in \{absent, slight, moderate, serious\}$

$AD \Rightarrow Alzheimer'sDisease \Rightarrow \forall x_3 \in \{absent, present\}$

N=> Is a function that counts the number of records in the database of instances that fulfil the query criteria.



X: Instance_number (Num)

Y: TOTAL (Num)

Colour: Cluster (Nom)

Select Instance

Plot: DATA-weka.filters.unsupervised.attribute.Remove-R1-3,5-76_clustered

Fig. 4. Clusters for cognitive deterioration generated with WEKA

Once the clusters for Cognitive Deterioration and Cognitive Deterioration in Living Creatures and Non-Living Creatures have been defined, each instance of the database has to be categorised according to the number of attributes that the patient has produced for each object. We have to make the symptom variables discrete, and for this we use the K-means algorithm, with as many clusters and states that we want the variables to have. Once the symptom variables are discrete, the database is analysed to obtain the conditional probabilities thus:

$$P\left(CATx_1 \mid DCOx_2\right) = \frac{N(CATx_1, DCox_2)}{N(DCOx_2)}$$

where

$$CAT \Rightarrow SemanticCategories \left\{ \begin{array}{c} Taxonomical \\ Types \\ Parts \\ Functional \\ Evaluative \\ Place/Habitat \\ Behaviour \\ Causes \\ Procedural \\ LyfeCycle \\ Others \end{array} \right\} \Rightarrow \forall x_1 \in$$
$$\{absent, slight, moderate, serious\}$$

$$DC \Rightarrow SemanticDeficit \left\{ \begin{array}{c} Apple \\ Dog \\ Pine \\ Car \\ Chair \\ Trousers \end{array} \right\} \Rightarrow \forall x_1 \in$$
$$\{absent, slight, moderate, serious\}$$

N=> Is a function that counts the number of records in the database of instances that fulfil the query criteria.

4 Influence Diagram

Influence Diagrams can be used as a Practical Clinical Guide to diagnose Alzheimer and to determine in a normative way which is the most appropriate action policy to take even for cases where it is not as evident and the doctor's clinical judgement cannot provide a better solution. They could also contribute to the application of more recent research studies, even explicitly and systematically combining different experts' opinions and the experimental data that are obtained. It should be borne in mind that in all the research process computer tools are being updated and upgraded and they are easily extensible For that reason, we have wanted to refer to the Influence Diagrams as a fundamental component for this research.

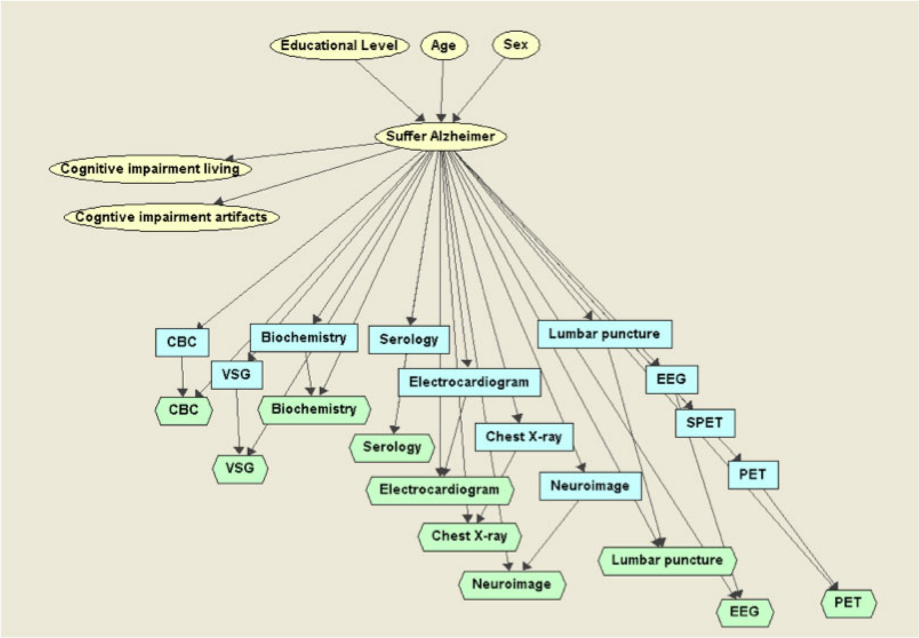


Fig. 5. Decision diagram for complementary explorations

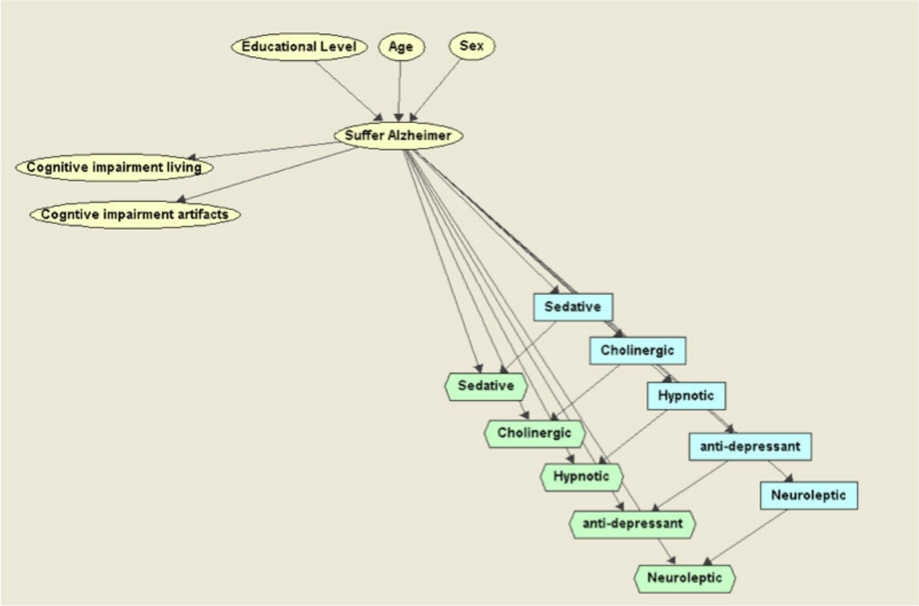


Fig. 6. Influence Diagram. Pharmacological treatments.

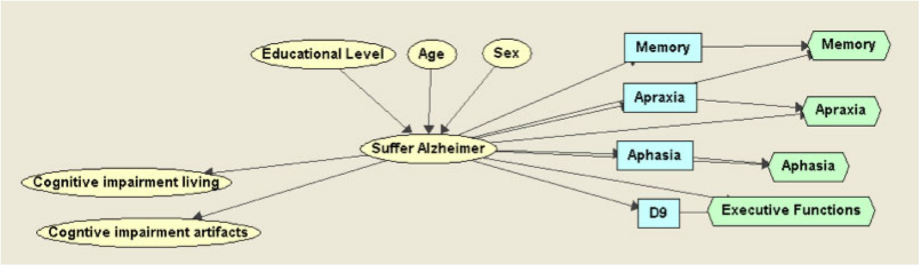


Fig. 7. Influence Diagram. Cognitive therapies.

This article proposes three influence diagrams: one (fig.5) to help determine whether complementary explorations must be done, another (fig.6) to help decide the most appropriate pharmacological treatment, and finally (fig.7), a decision diagram for cognitive treatment.

In the first Influence Diagram the following complementary explorations have been taken into account: CBC, VSG, Biochemical, Electrolytes, Hepatic function, Thyroid Function (T3, T4 and TSH), Vitamin (B12 and folic acid), Serology (Syphilis), Electrocardiogram, Neuroimage (CAT or Nuclear magnetic resonance), Lumbar Puncture, Electroencephalogram, SPECT (simple photon emission computed tomography) and/or PET (positron-emission tomography).

We have also constructed an Influence Diagram to analyse the maximum usefulness expected from each of the possible pharmacological treatments.

The last Influence Diagram evaluates the maximum usefulness expected in the application of cognitive therapies, i.e. psychological therapies based on the fundamentals of cognitive psychology.

5 Conclusions

This research work is in its initial stage, for that reason we have focused exclusively on the model and the decisions taken to design it. We believe that in the very near future we can obtain promising results, especially when we have a wider database of instances and more random instances in order to construct a realer quantitative model, which will enable us to do sufficiently reliable and assessable experiments. We are also convinced that Bayesian Networks can provide researchers with a powerful tool with great analytical capacity. Bayesian Networks and in particular the model that we present in this research work can include new variables, be they risk factors, symptoms or intermediate variables, and it is possible to analyse mathematically the impact that these variables can have on the diagnosis. This analytical capacity along with experts' epidemiological studies or subjective expert assessments could enable us to characterise the disease in its very initial stages. Furthermore, Influence Diagrams can establish action policies in a normative way, making it possible to apply and extend the most recent research studies in the Diagnosis of Alzheimer. Influence Diagrams could have great potential when applying and extending the research studies that are obtained.It should be highlighted that this research project opens a large multidisciplinary research field.

Acknowledgements

The authors are grateful to the CiCYT for financial aid on project TIN-2010-20845-C03-02.

References

1. Arias-Calleja, M.: Carmen: una herramienta de software librepara modelos gráficos probabilistas (2009)
2. Bottcher, S.G., Dethlefsen, C.: Learning bayesian networks with r. In: International Workshop on Distributed Statistical Computing, DSC 2003 (2003)
3. Cree, G.S., McRae, K.: Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). Journal of Experimental Psychology: General 132, 163–201 (2003)
4. Díez-Vegas, F.J.: Teoría probabilista de la decisión en medicina (2007)
5. Fernández-Galán, S., Díez-Vegas, F.J.: Modelling Dynamic Causal Interactions with BayesianNetworks: Temporal Noisy Gates (2000)
6. Kjaerulff, U.B., Madsen, A.L.: Bayesian Networks and Influence Diagrams
7. Lacave, C.: Explicación en Redes Bayesianas (2002)

8. McRae, K., Cree, G.S., Seidenberg, M.S., McNorman, C.: Semantic feature production norms for a large set of living and non living things. Behaviour Research Methods 37, 547–559 (2005)
9. Moreno, F.J., Peraita, H.: Análisis de la estructura conceptual de categories semánticas naturales y artificiales en una muestra de pacientes de alzheimer. Psicothema 18(3), 492–500 (2006)
10. Neapolitan, R.E.: Learning Bayesian Networks. Series in Artificial Intelligence. Prentice-Hall, Englewood Cliffs (2004)
11. Nielson, T.D.: Bayesian Networks and Decision Graphs (2007)
12. Peraita, H.: Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad del alzheimer. Technical report, Departamento De Psicología Básica 1. Facultad de Psicología. UNED (2009)
13. Peraita, H., Galeote, M.Á., González-Labra, M.J.: Deterioro dela memoria semántica en pacientes de alzheimer. Psicothema 11(4), 917–937 (1999)
14. Peraita, H., Grasso, L., Mardomingo, M.C.: Análisis preliminar de rasgos de definiciones de categorías semánticas del corpus lingüístico de sujetos sanos y con enfermedad de alzheimer, Technical report, Departamento de Psicología Básica 1. Facultad de Psicología. UNED (2009)
15. Valls-Pedret, C.: Diagnóstico precoz de la enfermedad de alzheimer: fase prodrómica y preclínica. Rev. Neurol. 51(8), 471–480 (2010)