Knowledge Discovery from Observational Data of Causal Relationship between Clinical Procedures and Alzheimer's Disease

Soumaya Yacout[*1], Alya Danish [*2] , Susie ElSaadany [◦3] , Jean-Pierre Kapongo [•4] , Suja Mani [◦5] ,James Gomes [•6]
[#] École Polytechnique de Montréal, Canada

*Université de Montréal
◦ Public Health Agency of Canada, Canada
• University of Ottawa, Canada
[1]soumaya.yacout@polymtl.ca;[2]Alya.danish@umontreal.ca; [3]susie.elsaadany@phac-aspc.gc.ca; [4]jptkapongo@yahoo.co.uk; [5]smani555@yahoo.ca;[6]James.Gomes@uottawa.ca

***Abstract*- In this paper, a knowledge discovery tool called Logical Analysis of Data is used to shed light on the causal relationship, if any, between three clinical procedures, namely blood transfusion, surgery and organ transplant, and Alzheimer's disease, which is thought to be a prion-type disease of protein misfolding, capable of spreading infectiously from human to human. The Logical Analysis of Data is a data-mining artificial intelligence technique that allows the classification of phenomena based on knowledge extraction and pattern recognition, without the reliance on prior hypotheses or any statistical analysis.By creating a database of clinical information obtained from a systematic review of the literature on the risk factors of Alzheimer's Disease, we were able to apply the Logical Analysis of Data to reveal the patterns distinguishing cases of AD that have undergone any of the three clinical procedures, and those cases that have not. Although several eye-opening patterns were revealed, results show that there is no evidence of relation between blood transfusion, surgery or organ transplant and the onset or development of Alzheimer's disease.**

Keywords-Data Mining; Logical Analysis of Data; Alzheimer's Disease; Blood Transfusion; Surgery; Organ Transplant; Pattern Recognition; Systematic Review

I. INTRODUCTION

Despite the primary importance of the identification of causal relationships between the development and onset of Alzheimer's disease (AD) and associated risk factors, modeling of the underlying disease mechanisms has been challenging[1]. For many public health problems, causal relationships are complicated, nonlinear and dynamic, which adds to the complexity of understanding, preventing and treating these diseases. Moreover, although a number of studies on the AD can be found, these studies are not necessarily of similar design, and the variability between them renders the application of rigorous statistical meta-analysis techniques almost impossible. In this paper, a knowledge discovery approach called Logical Analysis of Data (LAD) is introduced in order to study the relationships between clinical procedures, namely blood transfusion, cell, tissue and organ transplantation and surgeries, and the onset and development of AD. LAD is a data mining artificial intelligence approach that has previously been used in medical research (e.g., [2], [3]). It is a pattern recognition and classification approach that takes advantage of the latest developments in the field of information technology, namely the increase in the speed of computation and the storage and analysis of large volumes of data. The purpose of this paper is thus to introduce LAD in order to develop a better understanding of the effect or causality, if they exist, of clinical procedures on the risk of AD, or conversely, the effect of AD on clinical procedures.

Dementia diseases and their most common manifestation, Alzheimer's disease (AD), are on the rise across the globe. Approximately 35.6 million people are currently living with dementia, according to estimates by the World Health Organization, which projects that this number will double by 2030 and more than triple by 2050 [4]. In Canada, over 500,000 individuals are affected by AD, with 60,150 new cases reported each year among Canadians aged 65 and over [5, 6]. In the United States, the incidence and prevalence of AD has seen a dramatic increase by age found within all race/sex strata as estimated in the Cardiovascular Health Study (CHS) cohort[7].

Alzheimer's disease is a progressive, irreversible, neurodegenerative disease which begins with insidious deterioration of higher cognition and progresses to severe dementia, in which memory and thought processes become impaired, affecting intellectual and social skills to the point where daily life becomes difficult. A subtle deterioration of memory progressing to profound memory loss, loss of mental powers (the ability to think, understand, reason, learn, and solve problems), personality changes, and an increasing inability to carry out the activities of daily living are characteristics of AD [8]. Research has shown that AD is caused by complex interactions between a number of risk factors such as age and sex, genetic factors such as the presence of the APOE_4 allele or a previous family history of dementia and environmental and lifestyle factors, such as

DOI: 10.5963/PHF0201001

alcohol or tobacco use [9]. The role of many of these risk factors has been well characterized, but for others, there are still some uncertainties. One group of risk factors that is not fully understood is clinical procedures involving blood transfusion, cell, tissue and organ transplantation and surgeries.

AD shares a number of features in common with prion diseases, a heterogeneous group of conditions defined by brain accumulation of abnormal prion protein PrPsc and transmissibility. The five human phenotypes, based on clinical, pathological, biochemical, and genetic criteria, include: Creutzfeldt-Jakob disease (CJD), Gerstmann-Straussler--Scheinker syndrome (GSS), fatal familial and sporadic insomnia (FFI and sFI), kuru, and new variant Creutzfeldt-Jakob disease (nvCJD).Common features of these diseases include age requirement, mutations in an amyloidogenic protein, copper binding properties of the amyloid ogenic protein, evidence of free radical damage, the presence of polymorphisms that influence disease susceptibility, formation of amyloid plaques, and in some cases the presence of neurofibrillary pathology[10]. Current data suggest that for prion diseases, normal cellular PrP (PrPc) is converted into PrPsc through a process whereby alpha helical regions are refolded into beta sheets[10]. Mutant prion proteins convert normal prion proteins into misfolded, pathogenic forms, where the misfolded proteins accumulate together forming protein aggregates (amyloid forms). This activity forms the basis of infectivity. Prions are exceptional in that they are able to enter their hosts by natural portals and make their way from the gut to the brain, utilizing intermediate tissues for amplification. They are therefore transmissible by natural routes, mainly by ingestion. Mutant mammalian prions transmit spongiform encephalopathies, which are untreatable disorders. Spongiform encephalopathies are thought to be transmitted by blood transfusion, even prior to the clinical onset[11].

The ability of prion proteins to access self-templating amyloid forms (misfolded protein aggregation) confers phenotypic changes that can spread from individual to individual within or between species[12]. In fact, this ability is not unique to prion proteins. Several fatal neurodegenerative diseases are associated with the accumulation of self-templating amyloid forms of other types of proteins. For example, β amyloid (Aβ) and tau misfolds in AD, αsynuclein misfolds in Parkinson's disease, and Huntington misfolds in Huntington's disease. Neurodegenerative diseases (such as AD)then, and their associated amyloid forms, are thought to spread from cell to cell within the brains of afflicted individuals, thereby spreading the specific neurodegenerative phenotypes distinctive to the protein being converted to amyloid [12]. Amyloid A amyloidosis is a neurodegenerative disease which is thought to be transmitted between individuals by peripheral blood monocytes, highlighting a potential risk for human-to-human transmission through blood transfusions [11].Individual transmission of AD is said to occur through exposure to blood, cells, tissues or organs via blood transfusions, organ transplant, contaminated surgical tools or the use of drugs of human origin [12]. In this paper, we will attempt to shed light on this affirmation by posing the following question: Based on observational data, is there enough evidence to confirm that the latter statement is correct?

II. THE LOGICAL ANALYSIS OF DATA

A relatively new approach called the Logical Analysis of Data (LAD) for data mining and knowledge discovery will explore the cause-effect relationship between clinical procedures and AD. LAD is an inductive approach that does not start from any hypothesis. It is based solely on the exploitation of databases of observations, a technique that is now possible due to the tremendous advancement in the field of information technology and computers.LAD is an artificial intelligence technique that allows the classification of phenomena based on knowledge extraction and pattern recognition. The patterns found in the observational data are the building blocks of LAD. These patterns describe knowledge hidden in the database. LAD can recognize patterns distinguishing cases that have the disease from those that do not. A positive pattern is a combination of values of the attributes occurring together in only some of the observations for cases with the disease. These patterns have never occurred or were never identified in patients who do not have the disease. Alternatively, a negative pattern indicates values of the attributes occurring together in cases who do not have the disease and which were not identified in cases that do have the disease. As with many artificial intelligence techniques, LAD is applied in two consecutive phases, the learning or training phase, and the testing or theory formation phase, where part of the database is used to extract special patterns of phenomena and the rest of the database is used to test the accuracy of the previously learned knowledge. We note that LAD is a technique based on supervised learning; this means that the database contains the attributes' measurements and their corresponding classes or outputs. In this study, the output or class is the state of a case diagnosed as having AD (denoted by 1 or positive) or not having AD (denoted by 0 or negative). The state or class can also be the reception of a procedure (1 or positive) or no-procedure (0 or negative).

In this research, we examine the patterns that differentiate between cases diagnosed with AD and never having undergone any clinical procedure, and those diagnosed with AD and having undergone a procedure. A procedure is defined as any clinical intervention carried out to improve, maintain or assess the health of a patient. The procedures examined are: blood transfusion; cell, organ and tissue transplant; and surgery. A pattern is a combination of the values of attributes observed in cases of AD that have never undergone any clinical procedure, and have never been observed in cases that were diagnosed with AD and have undergone a procedure. In the following sections, we describe how a systematic review of the literature was performed to obtain the data input for LAD. We then detail how LAD was used to develop a better understanding of the effect of clinical procedures on AD, or if these diseases are related to the clinical procedures being studied. The following research questions are asked: Based on what is published in the literature, are there different patterns for cases having received the clinical procedures and those that did not? Are these patterns the same for cases diagnosed with AD before the procedure or those diagnosed after the procedure? And, are clinical procedures parts of these patterns?

DOI: 10.5963/PHF0201001

III. METHODS

A. Systematic Review

Since direct clinical data were not available, a review of the scientific literature was executed using specifically designed criteria, followed by a synthesis of the data into a database. This enabled us to analyse previous research by using LAD to extract the information embedded in the database and to search for patterns. A search strategy in publications between 1990 and 2012 was developed using the PubMed search engine, and after having optimized the search strategy, other databases such as Ovid Medline, Tox line, Cancer Letters, Scholars Portal, Embase and other sources of gray literature were sought for relevant scientific publications. We first identified 59363 publications. The work was then divided in three sections. First, articles were grouped that generally address the topic of AD. Secondly, identified publications were filtered by adding the independent predictors of risk of AD considered for this study, which include the following risk factors: age, gender, level of education, alcohol consumption, tobacco use, physical activity, frailty, cognitive capacity, presence of any of the following conditions: dementia, biological deficiencies (including hemoglobin, plasma hemocystine and polycythemia),anemia, diabetes, obesity, genetic deficiencies, hypertension, inflammation, hearing loss, presence of a specific diet, use of NSDAIDS (Ibuprofen, Aspirin, Ketoprofen, Sulindac, Naproxen, Etodolac, Flurbiprofen, Ketorolac, or Proxican) and previous trauma to the head. The presence of other medical conditions, namely amyloid deposition, Hyperhomocysteinemia, alcohol hepatopathy and orthotropic hepatic transplant and Parkinson's disease were also considered. The third step was to map therelevant scientific papers against each of three clinical procedures: blood transfusion, surgery and transplants, as well as the timing of these procedures as before or after the diagnosis of AD.Inclusion criteria were papers published in last 12 years, controlled case studies (trials) and English or French languages. Quality assessment of the publications was conducted using the Downs and Black criteria. The results of this systematic review were 4802 publications of which 108 were finally retained for inclusion in the study. These articles studied 168350 cases. The selected articles were organized in a database created in Ref Works and subsequently, Distiller was used for data extraction. The data were then recorded and tabulated in an Excel spreadsheet and was then analyzed using the cbm LAD program[13] to assess the relevancy of risk factors and clinical intervention on the onset of AD. Among the 108 publications, 19 publications discussed diseases that appear after some clinical interventions (mostly surgeries) such as: idiopathic nasal CSF leak correction, bile duct legation, general surgery, cardiac surgery, prostate & hernia surgery, biliar anastomosis and deep brain stimulation of the sub-thalamic nucleus. In this category, the case of AD was found in 13 cases, dementia in 6 and both diseases in 2 cases after the intervention. 33 publications found AD without any intervention, but the disease was induced by some risk factors. 3 articles addressed dementia without any surgical intervention, but the disease was induced by risk factors. 18 publications discuss both AD and Dementia, where individuals had never had any clinical intervention, but the diseases were prompted by some risk factors. A total of 71 research papers on risk factors prevailed in the pathology of AD and dementia. These research articles reported on the following risk factors: age (15 cases), alcohol (4cases), biological factors (7 cases), diabetes (6 cases), diet (7 cases), education (3 cases), frailty (3cases), gender (6 cases), genetics (24 cases), hypertension (11cases), inflammation (4 cases), obesity (5cases), tobacco (2 cases), anemia (1case), trauma to head (4cases), other medical conditions (10 cases). Each of 71 of these papers was reviewed and the required data were extracted according to a priori decided criteria. Information was extracted on the research questions, characteristics of the population that was studied, study design, methods used to collect the required information and analyses that were conducted. The clinical procedures that were considered eligible for inclusion in this review were any surgical procedure (35 cases), blood transfusion (9 cases) and cell, tissue or organ transplant (8 cases), prior to or after the diagnosis of Alzheimer's disease or dementia. 37 research papers were finally added through targeted search literature.

B. The Application of LAD

The Logical Analysis of Data (LAD) identifies interactions between risk factors (attributes) based on observational data, without any prior hypotheses or statistical assumptions. It is a combinatorics and optimization based data analysis method that depends mainly on the continuous development in the field of information technology, namely the speed of computation and the volume of data that can be analyzed. The basic idea of LAD is to combine a differentiation/ integration approach of a subspace of R^n, where n is the dimension of real vectors describing a dataset consisting of two disjoint sets O^+ and O^-. Typically each of the vectors appearing in the dataset corresponds to a patient or a case, the vectors in O^+ correspond to cases having a specific medical condition (e.g. AD), while those in O^- (the 'control' in medical language) do not have that condition. The components of the vectors called 'attributes' or 'features' or 'variables' $I_1, I_2, ...I_m$, represent the values of certain measurements, tests, histories (e.g. clinical procedures), general information about the patients, such as the presence or absence of certain symptoms. The power of LAD is in finding a family of homogeneous subsets P^+ (respectively P^-) of patterns, having a significant intersection with O^- (respectively O^-) but being disjoint from O^- (respectively O^+). These intervals are called 'positive (respectively negative) patterns'. Much information can be extracted from the subsets P^+ (respectively P^-). Among the most important information is the discovery of combinatorial 'biomarkers', i.e. the highly

DOI: 10.5963/PHF0201001

influential combinations of attributes, the 'blockers', i.e. combination of attributes inhibiting the medical condition, and the 'promoters', i.e. attributes that favor the medical condition. There are two major steps to LAD: Binarization of data, and the pattern generation procedure.

1) Data Binarization:

LAD was originally developed for the analysis of datasets whose attributes take only binary (0-1) values (e.g. male-female, diet-no diet, alcohol-no alcohol). Since many real-life applications include continuous variables (e.g. age), a 'binarization' method was proposed. The basic idea consists of the introduction of several attributes associated to each of the numerical variables. These attributes take the value "1" (respectively 0) if the numerical variable to which it is associated is above (respectively below) a certain threshold. Mathematically, many binarization techniques are suggested in the literature. The simplest technique consists of ranking the distinct values u that an attribute takes in a non-decreasing order. Then a cut-point α is inserted between each two consecutive values that belong to different sets of $O = O^+ \cup O^-$. The cut-point is calculated as the average of the two values. A binary attribute x is then formed from each cut-point such that:

$$x(u) = \begin{cases} 1 & if \quad u \geq \alpha \\ 0 & if \quad u < \alpha \end{cases}$$

The number of binary attributes that make up the binarized database at the end of the binarization process is equal to the number of cut-points generated for each continuous variable.

2) Pattern Generation:

A basic concept in the Logical Analysis of Data is that of generating patterns. It resembles the concept of rules that appears in various artificial intelligence techniques, but differs in that these rules are not imposed by anyone, they are rather found in the dataset. There are two types of patterns, positive and negative. By convention positive means having the specific condition, while negative means condition-free. Mathematically, the LAD approach is based on Boolean theory. A Boolean function $f(x_1, x_2, ...x_n)$ is a mapping $[0,1]^n \to [0,1]$, where $x_1, x_2, ...x_n$ for n>m are variables obtained by binarizing the attributes $I_1, I_2, ...I_m$, based on a binarization procedure, for example the one presented in this paper. A partially defined Boolean function (pdBf) is given by a set of ndimensional 0-1 vectors and is denoted by (O^+, O^-), where $O^+ \subseteq [0,1]^n$ is the subset of "true" (or "positive") vectors, and $O^- \subseteq [0,1]^n$ is the subset of "false" (or "negative") vectors. A Boolean function $f(x_1, x_2, ...x_n)$ is called an extension of a pdBf if:

$$f(X) = \left\{ \begin{matrix} 1 if\ X \in O^+ \\ 0 if\ X \in O^- \end{matrix} \right\}, X \equiv (x_1, x_2, ..., x_n).$$

A literal is either a binary variable x_i or its negation $\overline{x}_i$. A term C is a conjunction of distinct literals which does not contain both a variable and its negation. The degree of a term is the number of literals in it, for example (01001) is the term ($\overline{x}_1 x_2 \overline{x}_3 \overline{x}_4 x_5$) of degree five. A term C covers a vector X if C(X)=1. The Boolean sub cube of $[0,1]^n$, not necessarily included in $O^+ \cup O^-$, corresponding to the observations covered by a term C, is denoted by S(C), while $S(C) \cap O^+ \cup O^-$ is called coverage of C denoted by $COV(C)$.

A term C is called a positive or (negative) pure pattern p of a pdBf (O^+, O^-) if:

1. $C(X) = 0$ for every $X \in O^-$ ($respectively\ X \in O^+$) and

2. $C(X) = 1$ for at least one vector $X \in O^+ (X \in O^-)$

These two conditions mean that positive patterns $p_i \in P^+, i = 1, 2, ..., K$, where K is the number of positive patterns, are generated and P^+ is the set of these patterns and it is a vector of binary values covering at least one observation having the specific condition, that is in the positive group. This same pattern is not found in any observation in the 'control' group that is the negative group. The reciprocal is also true for the negative pattern $p_i \in P^-, i = 1, 2, ..., Q$, where Q is the number of negative patterns generated and P^- is the set of these patterns. This type of pattern is called pure. Although these patterns are

DOI: 10.5963/PHF0201001

specific to one class, positive or negative, they may be difficult to find and thus some observations will not be covered by any patterns at all. In order to cover every observation in the database by at least one pattern, we generate mixed patterns that cover mostly positive (or negative) observations and some negative (respectively positive) observations according to user defined parameters. These patterns, although they define less accurately the information hidden in the database, they allow the coverage of every observation at least once. Many techniques for the generation of patterns have been proposed (e.g. [14],[15]). In this paper we use a modified version of Ryoo's formulation for pattern generations.This formulation is shown in Equations (1) to (8). In that formulation, as the operation of pattern generation involves the generation of positive and negative patterns, a pattern of a certain class and its opposite are referred to by the notations $*$ and $\overline{*}$ respectively. The Boolean pattern vector $W(w_1, w_2, ..., w_{2n})$ has a dimension that is double the number n of binarized attributes. If $w_j = 1$ then the attribute x_j is included in the generated pattern and its literal is equal to 1, and if $w_{n+j} = 1$ then the attribute x_j is included in the pattern and its value is equal to 0. Obviously both cases cannot exist at the same time as shown in Constraint (3). Each observation o_i is associated with a Boolean vector

$$\min_{\mathbf{w},\mathbf{y},\mathbf{d}} \sum_{i \in S^*} y_i$$

$$s.t. \begin{cases} \sum_{j=1}^{2n} a_{i,j} w_j + n y_i \geq d \quad \forall i \in S^* & (1) \\ \sum_{j=1}^{2n} a_{i,j} w_j \leq d - 1 \quad \forall i \in \bar{S^*} & (2) \\ w_j + w_{n+j} \leq 1 \quad j = 1, 2, \ldots, n & (3) \\ \sum_{j=1}^{2n} w_j = d & (4) \\ 1 \leq d \leq n & (5) \\ \mathbf{w} \in \{0,1\}^{2n} & (6) \\ \mathbf{y} \in \{0,1\}^{N^*} & (7) \\ \sum_{j=1}^{2n} r_{k,j} w_j \leq d_k - 1 \quad \forall \mathbf{r}_k \in \mathbf{R} & (8) \end{cases}$$

$a_i(a_{i,1}, a_{i,2}, ..., a_{i,n}, ..., a_{i,2n})$ such that $a_{i,j} = 1, \ (j = 1, 2, ..., n)$ if x_j is in o_i, and $a_{i,j+n} = 1, \ (j = 1, 2, ..., n)$ if $\overline{x}_j$ is in o_i. Finally, $Y(y_1, y_2, ..., y_{N^*})$ is a Boolean vector whose number of elements N^* equals the number of observations in the binarized training set S . The elements y_i of Y are the 0-1 binary decision variables that minimize the objective function, such that $y_i = 0$ if the observation $i \in S^*$ is covered by the generated pattern and equal to 1 otherwise. **R** is the set of generated patterns at each step. At the beginning, this set is empty, then it is updated after every pattern generated. At that time the parameter a is called r. Thus the problem formulation seeks the patterns with maximum coverage. We note that this formulation is not a linear modeling of the fault detection problem, but rather a procedure to pattern generation. Constraint (8) prevents the MILP algorithm from finding the same pattern twice, where R is the set of generated patterns so far, and d is the degree of these patterns. Many characteristics of patterns have been found [16]. Two important characteristics of patterns that will be used in this research are the degree and the prevalence of a pattern. The degree is the number of attributes that constitute the pattern. Usually we are more interested in low degree patterns because they are more general and easy to interpret. High degree patterns are more data specific and indicate more complex cause-effect relationships. The prevalence is the proportion of observations in one class (positive or negative) that are covered by a pattern. The higher the prevalence, the more important is the pattern. In this paper, we focus our attention on answering the following questions: Are there different patterns for cases

DOI: 10.5963/PHF0201001

with AD that have received a clinical procedure and those that did not? And, are these patterns the same for cases diagnosed before the procedure and those diagnosed after the procedure? Once these patterns are found by LAD, the risk factors that constitute them are the 'biomarkers', the 'blockers' or the 'promoters', depending on whether the patterns are negative or positive, respectively. The data generated from the systematic review that is described in Section A is used as input to the software cbm LAD[13].A sample of this data is given in Table 1.

TABLE I SAMPLE OF THE INPUT DATA

AD diagnosed	Age	Alcohol	Anemia	Biological	Diabete	.....	Procedure after diagnosis
1	0	0	0	1	0		0
1	0	0	0	1	0		0
1	0	0	0	1	0		0
1	1	0	0	1	1		0
1	0	0	0	0	0		0
1	0	0	0	0	0		0
1	1	1	0	1	0		0
1	0	0	0	0	0		0
1	0	0	0	0	0		0
1	0	1	0	0	0		0
1	0	0	0	1	1		0

Twenty-five risk factors, including the clinical procedures done before and after diagnosis, are considered as summarized in Table 2. The following legend was used:

1=yes and 0=No;

Biological: Includes hemoglobin, plasma hemocystine and polycythemia;

Transplants: Cell, organ and tissue transplants;

Age: Less than 65 = 0 and more than or equal 65=1;

Gender: female=1, male=0;

Genetic: DNA damage, RNA degradation, mitochondrial dysfunction, Amyloid beta peptide high serum;

NSAIDs: Ibuprofen, Aspirin, Ketoprofen, Sulindac, Naproxen, Etodolac, Flurbiprofen. Ketorolac, Proxican;

Other: Includes lack of physical activity=1;

Transplants: Include cell, organ and tissue transplants; transplant=1 indicates presence of at least one of three transplants;

Type of surgery: idiopathic nasal CSF leak correction=1, bile duct legation=2, general surgery=3, cardiac surgery=4, prostate & hernia surgery=5, biliary anastomosis=6, deep brain stimulation of the subthalamic nucleus=7, neural transplant=8, hydrocephalus surgery=9, corneal transplant=10, none=0

TABLE II DATA COLLECTION FORMAT USED TO COLLECT DATA FOR LAD PROCEDURE

Diseases or Classes			
AD diagnosed	DEMENTIA diagnosed	AD and Dementia	AD or Dementia
Risk Factors			
1.Age, 2.Alcohol, 3.Anemia, 4.Biological, 5.Diabete, 6.Diet, 7.Hiv status, 8.Frailty, 9.Gender, 10.Genetic, 11.Hypertension, 12.Inflammation, 13.NSAIDs, 14.Obesity, 15.Polycythemia, 16.Tobacco, Exposure to the, 17.Trauma to head			
Procedures			
19.Blood transfusion	20. Surgery	21. Transplants (cell, tissue, organ)	22.Type of surgery
Medical conditions & procedures before or after diagnosis			
23. Other medical conditions	24. Procedure before diagnosis	25. Procedure after diagnosis	

In order to answer the research questions, we searched for homogeneous subsets P^+ (positive) (respectively P^- (negative)) of patterns, having a significant intersection with O^+ (respectively O^-) but being disjoint from O^- (respectively O^+). This was done in five steps where:

DOI: 10.5963/PHF0201001

1. O^+ is the set of cases having AD and O^- is the set of cases that do not have AD.

2. O^+ is the set of cases having AD and O^- is the set of cases that do not have AD, and both sets consist only of cases which had undergone at least one clinical procedure.

3. O^+ is the set of cases that had undergone at least one clinical procedure and O^- is the set of cases that did not undergo any procedure.

4. O^+ is the set of cases that had undergone at least one clinical procedure and O^- is the set of cases that had not undergone any procedure, and in both sets all the cases have AD.

5. O^+ is the set of cases that received a procedure before diagnosis of AD and O^- is the set of cases that received a procedure after the diagnosis of AD. Obviously this means that all cases received a procedure. In the two sets, all cases have AD.

Positive and negative patterns were generated at each of the five steps. A sample of these patterns is given in Table 3.

TABLE III POSITIVE AND NEGATIVE PATTERNS THAT ARE OBTAINED IN EACH OF THE FIVE STEPS OF THE ANALYSIS

Step	Pattern	Class	Degree /number of attributes
Step 1: AD vs No-AD	P1: Age less than 65 Anemia, No Diabetes , No Frailty, No NSAIDs, No Obesity, No Polycythemia, No Others, No Trauma to head, No Blood transfusion, No Surgery No Transplant No	positive	13
	P2: Procedure after diagnosis of Dementia, Yes	Positive	1
	P3: Hypertension, Yes	Positive	1
	P4: Inflammation, Yes	Positive	1
	P5: Others, Yes	Positive	1
Step 2: AD vs No AD, all cases had at least one procedure	P6: Dementia, Yes Education, No Genetics, No	negative	3
	P7: Biological, No Procedure before diagnosis, Yes	positive	2
Step 3: Procedure vs No procedure	P8: Diabetes, Yes	Negative	1
	P9: Obesity, yes	Negative	1
	P10: Anemia, Yes	Negative	1
Step 4: Procedure vs No procedure, all cases have AD	P11: Trauma to head, Yes	Negative	1
	P12: Others, Yes	Negative	1

III. RESULTS

In each of these steps listed above, LAD generated a number of patterns, for example in the first step LAD generated 12 patterns. The first is an example of a high degree pattern, which has a high prevalence of 40%. This pattern is given as an example in Table 3. The others are examples of low degree patterns, mostly 1 degree, and are easy to interpret as was explained in Section II.B. All these patterns are pure, that is they cover some observations in one class and never in the other. In Step 1, P1 states that some of the cases having AD are less than 65 and they have no anemia, diabetes, frailty, NSAIDs, obesity, polycythemia, other medical conditions (as defined in Section A), trauma to the head, blood transfusion, surgery or transplant, but none of the cases that do not have AD has this pattern. P2 states that none of the cases that did not have AD had a clinical procedure after having been diagnosed with dementia. P3 states that none of the cases that do not have AD has hypertension. P4 states that none of the cases that do not have AD has inflammation. P5 states that none of the cases that do

DOI: 10.5963/PHF0201001

not have AD has any other medical condition. In Step 2, P6 states that the cases of AD who had a procedure were never diagnosed with dementia and had a level of education less than secondary school, referred to as 'no education' and 'no genetic diseases'. P7 states that while some of the cases that had AD diagnosed after a clinical procedure (namely surgery) had biological deficiencies; none of the cases that do not have AD and had a clinical procedure had any of the said deficiencies. In Step 3, P8 states that cases that have undergone a clinical procedure do not have diabetes

and that cases that have diabetes did not undergo a clinical procedure. P8 states that cases that had a procedure do not have obesity. P10 states that cases that had a procedure do not have anemia. In Step 4, in addition to P8, P9, and P10 of Step 3, P11 states that cases that had a clinical procedure and have AD have never had trauma to the head; P12 states that cases that had a procedure and AD have never had other medical conditions as defined in Section A. In Step 5, P13 and P14 state that cases that have AD have never had a blood transfusion or a transplant before the diagnosis. P15 states that for cases that have AD, the procedure before diagnosis is always a surgery of Types1, 2, 3, 4 or 5 only, and from this fact we observe that procedures after diagnosis of AD are always either a surgery Type 8 (neural transplant), a blood transfusion or an organ transplant.

IV. DISCUSSION

Based on the results obtained, LAD data exploitation and knowledge discovery leads to the following observations:

1. In Step 1 and from Pattern P1we discover that the cases of No AD have at least one of the attributes mentioned in P1 as risk factors, but they do not have each of the risk factors of hypertension, inflammation.

2. In Step 2, and from P6 and P7, we observe that patterns that differentiate between cases that have AD and received a procedure before the diagnosis and those that do not have AD and received a procedure (namely surgery). First, from P6 we found that some cases that were diagnosed with AD after a procedure also had dementia, 'no education' and no genetic conditions, while none of the cases of no AD has this pattern. Second, some of the cases that were diagnosed with AD after a procedure have biological deficiencies; while none of those has undergone a procedure and are free of AD have biological deficiencies.

3. There are different patterns for cases with AD that have undergone a clinical procedure and those that have not. For example, we gave patterns P8 to P12 which demonstrate that while some cases of AD that did not undergo a procedure have diabetes, obesity, anemia, trauma to head, and other medical conditions, none of the cases that had undergone a procedure suffers from any of these risk factors.

4. The patterns are not the same for cases diagnosed before the procedure and those diagnosed after the procedure. For example, in Table 3 we gave patterns that state that cases that have AD have never had a blood transfusion or a transplant before the diagnosis. For the cases that have AD, the procedure before diagnosis is always a surgery ofTypes1, 2, 3, 4 or 5 only, and the procedures after diagnosis of AD are always either a surgery Type 8 (neural transplant), a blood transfusion or a transplant. A strong consequence of these patterns is that there is no evidence to support the statement that blood transfusion and transplant lead to AD, since according the patterns; these were received only after diagnosis.

5. Clinical procedures are part of many of the observed patterns, for example, as stated in Point 4 above, surgeries of Types1, 2, 3, 4 and 5 are procedures associated with AD before diagnosis of the disease, while surgery Type 8, blood transfusion and transplant are procedures associated with AD after diagnosis. This is a clear example of biomarkers. An example of a blocker is a procedure vis à vis dementia, since no procedure was found to be undergone after diagnosis of dementia.

6. There is no evidence that AD leads to clinical procedures since none of the patterns found in Step 3 have AD as an attribute.

Finally it is worth noting that some of the obtained results are based on a very small number of observations. While they are true for the databases used in this study, we are hoping to apply the LAD approach to bigger databases in order to obtain more general results.

V. CONCLUSIONS

This paper presents an application of a relatively new approach called the Logical Analysis of data for data mining and knowledge discovery concerning the cause-effect relationship between clinical procedures and Alzheimer's disease. There are now numerous databases in the fields of clinical and public health that may be exploited, and LAD offers an efficient technique for this exploitation when traditional statistical analysis may fail for different reasons. An example which is relevant to this paper's study is the statistical technique called meta-analysis. This analysis is a quantitative synthesis of data obtained from several studies, where results are pooled, thus increasing sample size and the power to study effects of interest. However, these studies must be of similar design; only minor variability between the studies can be tolerated. The objectives of these studies, their clinical variables and outcomes must be precisely defined; the selection of studies and identification of bias must be rigorous; evaluation of heterogeneity among studies is paramount; data analysis techniques must be carefully designed and the use of sensitivity analysis is essential to ensure validity of a meta-analysis [14]. A successful meta-analysis also involves the

DOI: 10.5963/PHF0201001

identification and design of the appropriate statistical approach depending on the nature of the quantitative data. There are thus numerous caveats to the use of meta-analysis, which must be used judiciously in order to ensure the validity of its results. In contrast, LAD employs computational methods not usually used in traditional statistics. It is based on database technology and machine learning. It involves more "structureless" or "ad-hoc" analysis of datasets, which is antithetical to statistics. But it is this type of analysis of enormous datasets that can reveal unsuspected but valuable structures within them [15]. LAD deals with much wider data sets than statistics, and for which traditional statistical algorithms may be too slow. In such cases, the choice of model, a central endeavor in statistical work, is not obvious or may be impossible. LAD is not about fitting a model, but relies more heavily on pattern detection and can thus be useful in finding hidden relations. The central process of LAD is essentially exploratory, rather than "confirmatory", and contrasts with descriptive statistics in the sheer size of the data sets for which traditional statistics may fail[15]. Traditional statistics is mainly concerned with numerical data, whereas LAD deals with multi-dimensional and mixed forms of data at once: logical data, patterns composed of conjunctive and disjunctive combinations of elements and higher order structures [15]. Finally, inferential statistics such as hypothesis testing, is used when a scientific experiment is conducted, that is, the same sample of data is studied under different situations. The results of scientific experiment are then extrapolated to the research population. LAD allows for the examination of the entire dataset, which is usually big enough to allow us to consider that the entire research population is analyzed, rather than a sample from which inferences are then made to the larger research population. In this case, the use of statistical significance and hypothesis testing becomes obsolete [15].

In conclusion, the advantage of using LAD for data mining in public health can be summarized as follows:

1. LAD is not based on statistical analysis or prior hypothesis. Consequently, it does not assume that the data belong to a specific statistical distribution. The approach therefore does not require statistical analysis of data prior to or after its use. Unlike statistically based techniques, correlations and dependence between attributes or variables do not have any effect on LAD's performance. LAD can handle the interdependence between attributes, and moreover, it gives physical explanation to it, and to the interaction between multiple attributes.

2. LAD automatically extracts information from the generated patterns based on the observational data and, accordingly, classifies the cases into one class set based on the patterns generated.

3. The output of LAD can be traced back to the specific attributes that resulted in the categorization of the case into a certain class. The interpretability of all the results and of all the steps that lead to a specific output is clear. The patterns that are generated have physical meanings that are relevant to the user of LAD. They identify blockers, promoters and biomarkers.

4. LAD is a powerful knowledge management tool that automates and conserves knowledge. The more databases that are available, the more the use of LAD becomes relevant.

5. The approach of LAD in detecting and interpreting phenomena is based solely on the computational power of computers, specifically the speed of computation and the power of storage and treatment of a large volume of data. Thus, the continuous advancements in the field of information technology offer a strong leverage for the expansion of this approach.

We conclude by emphasizing that while the LAD approach has been proven to be mathematically rigorous, the results depend solely on the quality and the volume of the database. Although LAD can avoid many errors that may be present in the database [17,18], the accuracy of the output depends on the accuracy of the input. Finally, ideally the database is expected to represent the whole population of interest or at least enough historical experience in order to be worthy of analysis. If not, then the results can still be taken as knowledge discovered (eye opener), but not necessarily the truth, the whole truth and nothing but the truth.

REFERENCES

[1] Zaven S. Khachaturian (published online, 2006), Epilogue: Toward a Comprehensive Theory of Alzheimer's Disease-Challenges, Caveats, and Parameters. Annals of the NewYork Academy of Sciences. 924, Dec. 2000: 184-193.

[2] SorinAlexe, Eugene Blackstone, Peter Hammer, HemantIshwaran, Michael S. Lauer, Claire E. PothierSnader, "Coronary Risk Prediction by Logical Analysis of Data", Annal of Operations Research, 119, 2003, 15-42.

[3] Sacha D. Abramson, Gabriela Alexe, Peter L. Hammer, Joachime Kohn, A Computational approach to predicting cell growth on polymeric biomaterials. Wiley InterScience (www. interscience.wiley.com) DOI: 10.1002#jbm.a.30266.2005,116-124.

[4] Dementia cases set to triple by 2050 but still largely ignored. World Health Organization News Releases, 2012, http://www.who.int/mediacentre/news/releases/2012/dementia_20120411/en/index.html.

[5] Information about Alzheimer's and related dementias, 2012,http://www.cihr-irsc.gc.ca/e/45554.html.

[6] Ian McDowell, "The incidence of Dementia in Canada". Neurology. 55 (1): 66-73, 2000.

[7] A. L. Fitzpatrick, L. H. Kuller, D. G. Ives, O. L. Lopez, W. Jagust, J. C. S. Breitner, B. Jones, C. Lyketsos, C. Dulberg, "Incidence and Prevalence of Dementia in the Cardiovascular Health Study". Journal American Geriatrics Society 52:195–204, 2004.

[8] http://www.mdguidelines.com,Medical Disability Advisor.

[9] Christiane Reitz, Ming-Xin Tang, Nicole Schupf, Jennifer J. Manly,Richard Mayeux, Jose´ A. Luchsinger,"A Summary Risk Score for the Prediction of Alzheimer Disease in Elderly Persons", Arch Neurol.67 (7):835-841, 2010.

DOI: 10.5963/PHF0201001

[10] Rudy J. Castellani, George Perry and Mark A. Smith, “Prion disease and Alzheimer’s disease: pathogenic overlap”, ActaNeurobiolExp, 64: 11-17,2004

[11] Sponarova, J., Nystrom, S. N., Westermark, G. T. “AA-amyloidosis can be transferred by peripheral blood monocytes”. PLoS One 3, e3308,2008.

[12] M. Cushman, B.S. Johnson, O.D. King, A.D. Gitler, J. Shorter (2010), “Prion-like disorders: blurring the divide between transmissibility and infectivity”, Journal of Cell Science 123, 1191-1201, 2010.

[13] Soumaya Yacout, David Salamanca, Mohamed-Ali Mortada. Artificial intelligence tool for engineering applications and condition based maintenance. ÉcolePolytechnique de Montréal. Department of mathematics and Industrial engineering, 2009.

[14] Walker E. Meta-analysis: Its strengths and limitations. Cleveland Clinic Journal of Medicine, 2008. 75(6): 431-439.

[15] Hand DJ. (1999) Statistics and data mining: Intersecting disciplines. SIGKDD Explorations 1(1): 16-19.

[16] EndreBoros, Peter Hammer, Toshihide Ibaraki, Alexander Kogan, Eddy Mayoraz, IlyaMuchnik,”An implementation of Logical Analysis of Data”. IEEE Transactions on knowledge and Data Engineering, VOL. 12, NO. 2, 2000, pp. 292-305.

[17] Hong SeoRyoo, In-Yong Jang,“MILP approach to pattern generation in logical analysis of data”.Discrete Applied Mathematics. 157, 2009, 749-761.

[18] Gabriela Alexe, SorinAlexe, Peter Hammer, Alexander Kogan,“Comprehensive vs. Comprehensible classifiers in logical analysis of data”.Discrete Applied Mathematics, 156, 200, 870-882.

[19] Abderrazak Bennane, “Traitement des donnéesmanquantes pour cbmLAD ‘, M. Sc. Thesis, ÉcolePolytechnique de Montréal, Québec, Canada, 2010.

[20] Munevver Mine Subasi, ErosySubasi, Martin Anthony, Peter Hammer, “A new imputation method for incomplete binary data”, Discrete Applied Mathematics, vol. 159, 2011, 1040-1047.

DOI: 10.5963/PHF0201001