Identifying Combinatorial Biomarkers by Association Rule Mining in the CAMD Alzheimer's Database

Balázs Szalkai[a], Vince K. Grolmusz[c], Vince I. Grolmusz[a,b,*], Coalition Against Major Diseases[**]

[a] *Protein Information Technology Group, Eötvös University, H-1117 Budapest, Hungary*
[b] *Uratim Ltd., H-1118 Budapest, Hungary*
[c] *2nd Department of Internal Medicine, Semmelweis University, Budapest, Hungary.*

Abstract

The concept of *combinatorial biomarkers* was conceived around 2010: it was noticed that simple biomarkers are often inadequate for recognizing and characterizing complex diseases. It was proposed that biomarker-combinations should rather be considered and looked for. Here we present an algorithmic search method for complex biomarkers which may predict or indicate Alzheimer's disease (AD) and other kinds of dementia. In addition to commonly used statistical methods, we applied adequatly modified data mining techniques, namely association rule mining, that is capable to uncover implication-like logical schemes with quality scoring. The existing algorithms were modified to adopt the special needs of automatic combinatorial biomarker discovery: our DWARF program is capable finding multi-factor relevant association rules automatically. We applied the new DWARF program for a database of the Tucson, Arizona based Critical Path Institute CAMD (Coalition Against Major Diseases) AD database. The database contains the detailed laboratory- and cognitive test-data of more than 6000 patients from the placebo-arm of multi-million dollar clinical trials of large pharmaceutical companies, and consequently, the data is much more reliable than numerous other databases for dementia, derived from moderately

*Corresponding author

**Data used in the preparation of this article were obtained from the Coalition Against Major Diseases database (CAMD). As such, the investigators within CAMD contributed to the design and implementation of the CAMD database and/or provided data, but did not participate in the analysis of the data or the writing of this report.

funded research projects of probably looser standards. Some of the results reinforce known findings, therefore validate the method itself, while others can enlighten still unknown relations and biomarkers of dementia. We need to add that our goal was to find new biomarkers for Alzheimer's disease, but the database mostly contained cognitive test results that imply only the presence of dementia, and not necessarily Alzheimer's disease itself. The source code of the new DWARF program is publicly available in the supporting on-line information.

---

1. Introduction

Dementia is presently a major problem of high-income countries and also an increasing concern of low income nations worldwide. It is sporadic before age 60, but is doubled by every five years of age thereafter [1, 2]. About 40 percent of the population over 90 is affected, and up to 20 percent of population between 75 and 84 suffers from this condition [3, 4]. The most common cause of dementia is Alzheimer's disease (AD). The earliest symptoms of AD include memory problems; disorientation to time or place; and difficulty with calculation, language, concentration and judgment. As the disease evolves, patients may have severe behavioural abnormalities and may even become psychotic. In the final stages of the disease the sufferers are incapable of self-care and become bed-bound, even for years or even decades.

The causes and mechanisms of AD are not yet fully clarified. The underlying pathologic abnormalities include:

- Reduced levels in neurotransmitters, behindering inter-neuronal communication;
- Beta-amyloid plaques in and around synapses;
- A modified form of tau-protein shows accumulations in the neurons (neurofibrillary tangles).

Plaques and tangles mostly develop in brain areas vital for intellectual functions.

The diagnosis of AD in the great majority of the cases is done by clinical criteria, using standardized questionnaires [5]. Generally accepted evidences show that more than 20 years before those clinical signs the neuropathologic

damage begins [6], and by the time it is diagnosed, a large part of the neurons are already irreversibly lost.

In the last years, by the combination of the analysis of cerebrospinal fluid, clinical signs and neuroimaging techniques a quite reliable diagnostic method emerged [7]. The method, however, is prohibitively expensive, it is not an early warning-type biomarker, and does not seem to be applicable for wide-scale screening of the senior population.

Very recently, using the combination of usual clinical laboratory data, cognitive impairment questionnaires and blood-based proteomics assays was reported to reliably diagnose AD, without neuroimaging or cerebrospinal fluid assays [8, 9]. However, early warning biomarkers are still need to be found.

The final goal of ours is finding new combinatorial biomarkers for Alzheimer's disease. In this paper we report our results that may be used to reach this final goal; but presently we are able to show only that certain sets of laboratory data may make the dementia (and not the AD) more probable, and certain other sets may make the dementia less probable.

There are several large databases of Alzheimer's disease available for researchers. The quality of their data obviously depends on the methodology of the research that produced the database in question. Perhaps the most well-organized, strictly overseen and rigorously documented experiments are conducted by the order of large pharmaceutical companies in hospitals and clinics in phase 1, 2 and 3 drug trials. Unfortunately, the detailed results of those trials are seldom published (especially those that correspond to unsuccessful drug trials) since they are owned by the companies that ordered the trials.

The Tucson, Arizona-based Critical Path Institute made available in their Alzheimer's disease and Parkinson's disease database the results of the placebo-arm of numerous multi-million dollar clinical trials, conducted by the order of large pharmacological companies [10, 11, 12]. The data of the placebo-line of the trials does not contain proprietary information concerning the effects of the novel drugs under trial, but it does contain reliable, well-organized laboratory and cognitive test-data, presumably in much higher quality than other, larger, but perhaps less strictly conducted and controlled studies for AD.

Data used in the preparation of this article were obtained from the Coalition Against Major Diseases (CAMD) database [10]. In 2008, Critical Path Institute, in collaboration with the Engelberg Center for Health Care Reform

at the Brookings Institution, formed the Coalition Against Major Diseases (CAMD). The Coalition brings together patient groups, biopharmaceutical companies, and scientists from academia, the U.S. Food and Drug Administration (FDA), the European Medicines Agency (EMA), the National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute on Aging (NIA). The Coalition Against Major Diseases (CAMD) includes over 200 scientists from member and non-member organizations. The data available in the CAMD database has been volunteered by CAMD member companies and non-member organizations.

In contrast with more conservative statistical methods, we applied data mining techniques for the data analysis and combinatorial biomarker search. Data mining, as was defined in [13], is the analysis of large observational sets of data for finding new, still unsuspected relations with novel, usually high-throughput methods. Frequently, data mining uses large data sets collected for some other uses than the data mining analysis [13].

Simple biomarkers (e.g., the high glucose level in diabetes) show a physiological condition, related to the appearance, or the status of a disease. The concept of combinatorial biomarkers appeared around 2010, and numerous authors simply use the term in the following sense: If - say - the high concentration of all of the molecules A, B and C characterizes well a certain condition X (and the high concentration of any subset of the set A,B,C would not), then they say that A,B,C is a combinatorial biomarker of condition X [14]. In [9], by applying proteomics assays, a 30-protein set was identified as combinatorial biomarker of AD.

We intend to discover more involved combinatorial biomarkers, that may contain clinical laboratory data and psychiatric test data, and we count not only on the positive (i.e., high concentration or appearance of a certain value), but also their lack, or low concentration. We start with frequent item set analysis, then with association rule mining [13], with a new methodology, that discover more complex, combinatorial biomarkers only if they have stronger implications than the simpler biomarkers.

Therefore, our DWARF program will not produce artificially complex biomarkers: the more complex is the new biomarker the more valid is the new implication.

1.1. Association Rule Mining

Our research group was among the firsts applying association rule mining in molecular biology [15]. Recently, association rule mining gains applications

in drug discovery [16], in the design of clinical trials [17], and most recently, also in image analysis in Alzheimer's research [18].

Association rule mining is a field of data mining [13], developed by marketing experts for discovering implication-like rules in uncovering and customer behavior [19], without*a priori* assumptions on this behaviour. We intend to apply the method for laboratory and cognitive test data from the CAMD database [10]. We analyzed how the presence/absence/severity of cognitive impairment could be detected from combinations of known biomarkers, demographic information, measurements of vital signs. As an example, consider this expression:

$$sodium = high \wedge (protein = high \vee age \geq 60) \implies mmse\_total \leq 15 \quad (1)$$

This rule states that if blood sodium is high, AND urine protein is high OR age is at least 60, then the total MMSE (Mini Mental State Examination) score will be at most 15 out of 30 (i.e.,$\wedge$ stands for the logical AND and $\vee$ stands for the logical OR). Let us call the left-hand side (abbreviated by LHS) of the expression a combinatorial marker of the right-hand side (abbreviated by RHS). Thus the statement above can be reformulated as follows: high serum sodium combined with either high urine protein or age of at least 60 is a marker of a total MMSE score less than or equal to 15.

We considered all the possible logical expressions according to a given pattern, and assigned numerical values to them that indicated the reliability and validity of the logical rules. Then we filtered and sorted the vast amount of possible rules according to these numerical criteria, and selected the best ones. We changed a simpler rule to a more complex rule only if the more complex rule has higher reliability or validity than the simpler rule (see the next section for the exact definitions).

2. Methods

Our data source, which will be referred to as CAMD from now on [10], was provided by the Coalition Against Major Diseases, and consisted of the placebo arm of several drug trials. Over 6000 subjects participated in these trials including demented and not demented people of various age, sex, race and ethnics (see Table 4) for basic statistics). Standard laboratory data (including 300 different values in blood or urine, at different visit days) were

collected for the subjects altugether, though each person was tested for only about 30 different values. The cognitive and psychological status of the subjects was measured at different times by standardized questionnaires ADAS-COG, ADCS-ADL, MMSE, NPI and SIB. In addition, some genetic tests were performed, e.g., ApoE and MTHFR genotypes were recorded. Vital sign measurements (BP, pulse rate, respiratory rate and body temperature) were also taken. Results concerning this dataset will be described in greater detail below.

We transformed this large dataset into a conveniently processable form. The CAMD database contained several rows describing one person and these were scattered between multiple data tables. So we collected the essential data from CAMD into one single table: this simplified table contained only one row for each subject.

If a subject was tested on different visit days, then we took the average of these test results. The resulting main tables for CAMD consisted of around 170 columns (record fields) and 6000 rows (entries).

Our main method of processing these two resulting tables was association rule mining. First, we took a given pattern like $\square \wedge (\square \vee \square) \implies \square$. Notice that here the LHS (Left Hand Side) is in conjunctive normal form (multiple OR clauses ANDed together). This pattern can be described as "1 2", as the first OR clause has one sub-clause and the second one has two. This pattern matches all statements of the following kind: "if property A is present and property B or property C is present, then property D is present".

Since we are interested in implication-like association rules that indicate factors implying normal or demented mental state, we made restrictions on which data columns can occur on the LHS (Left Hand Side) and the RHS (Right Hand Side). Laboratory data, sex, race and ethnics were allowed on the on the LHS, and columns directly indicating mental status on the RHS. Then we gave numerical constraints on the "goodness" of a rule – thus introducing an ordering on the rules. Finally we tried to fill in all the void boxes in all possible ways to find the best rules.

If done without any optimization, this process would have yielded a vast amount of different rules that needed to be evaluated "by hand". Even just enumerating all the possible matches to this pattern would have required enormous computational resources. Consequently, we needed to make the computation feasible: we used a branch-and-bound approach similar to the Apriori Algorithm [13]: if certain values for the first two boxes made a rule fail our constraints – regardless of what would be written in the third box

–, then we threw out the rule and did not bother checking all the possible values for the third box. (An analogue could be cutting a tree in a clever way: one does not bother removing all the little twigs one by one, but rather cuts the trunk.) This technique saved us considerable computational time.

The association rule mining was done with our own program written in programming language D, named DWARF (D-written Association Rule Finder). The source code of DWARF and its documentation can be found in the supporting on-line material. D is a relatively new programming language offering both safety (through garbage-collection) and performance (it compiles to native code), that's why it was chosen.

We calculated various standard numerical values for all association rules, which would indicate their validity. First, we defined the *universe* of a rule: this is the set of the database rows where all columns present in the rule have a known value. For example, as we mentioned before, not all subjects were tested for everything, so our database contained a large amount of N/A entries. For testing the validity of a rule, only those rows could be taken into account, where there is no N/A written to any of the columns participating in the rule.

For evaluating the validity of a rule, we continued to work with only its universe and ignored all other rows in the database. Next, we calculated the *LHS support*, *RHS support* and *support* of a rule. The *LHS support* is the number of the rows where the LHS is true, the *RHS support* is the number of the rows where the RHS is true, and the *support* is the number of the rows where both the LHS and the RHS are true.

Then, we calculated the *confidence*, *lift*, *leverage* and χ^2-*statistics* for a rule. The *confidence* is defined as the conditional probability of the RHS, assuming that the LHS is true. In our example, confidence describes the chance having a low MMSE score, if one has high serum sodium combined with high urine protein or age at least 60. The *lift* shows how many times the presence of the LHS increases the probability of RHS. Generally it indicates how big a risk factor the LHS is – though it is not certain that the LHS *causes* the RHS, but they both may be just consequences of a phenomenon in the background [13].

The *leverage* is the difference between the observed probability of both the LHS and RHS being true, and the estimated probability we get by assuming that the LHS and RHS are independent events. Therefore, it indicates some dependency between the LHS and the RHS. Finally, the χ^2-*statistic* is a well-known measure of the estimated dependence of the indicator variables of the

LHS and RHS. The χ^2-statistic is greater than 3.84 if and only if the p value is less than 0.05.

The following table formalizes some of the above definitions. Here P denotes the probability measure, $P(A|B)$ denotes the conditional probaility of event A on condition B:

$$\begin{aligned} \text{Confidence} &= P(RHS|LHS) \\ \text{Lift} &= \frac{P(RHS|LHS)}{P(RHS)} \\ \text{Leverage} &= P(RHS \wedge LHS) - P(RHS)P(LHS) \end{aligned}$$

For the CAMD database the minimum acceptable values were set as follows: universe = 600, support = 65, confidence = 0.5, lift = 1.2, $\chi^2 = 3.84$. In particular, we recorded rules on data that were measured on at least 600 subjects. We defined the *goodness* of a rule to be equal to its lift.

Therefore we listed association rules of lift at least 1.2, i.e., only those rules were listed where the LHS increased the probaility of RHS with at least 20

As one of the most significant novelty in our approach, we filtered out rules which are too complicated: The DWARF program threw out elementary clauses from the LHS if the overall goodness (i.e., the lift) of the rule did not decrease by more than 2%, then threw out the whole rule if its numerical values dropped below our constraints during the simplification process. In other words, we sacrificed some of the lift for simplicity.

The program was run on a 16-core machine, so we divided the job into 20 parts and made a shell script spawn 20 simultaneous instances of the program. We chose a number bigger than the number of cores because the branch-and-bound technique makes runtime rather unpredictable, so some threads may finish earlier than others. Each thread looked for the 2000 best rules of its job slice. Then the results were merged into one text file.

Having listed the best rules, we also tried to determine whether the elementary clauses (like $lb\_ast = h$, $lb\_folate = l$, etc.) have positive or negative effect on mental state. Therefore we counted their appearances on LHS, and classified these occurrences by the nature of the RHS: does it indicate normal cognition or rather dementia? We counted how many times an elementary

clause occurred on the LHS of a rule when the RHS indicated a positive mental state, and how many times it occurred in rules where the RHS showed a negative state. Thus, in addition to mining rules whose LHS could probably serve as good combinatorial risk factor of dementia, we estimated the contribution of the *individual* clauses, for example "protein=*high*" to the onset of cognitive impairment.

For an elementary clause, *Positive score* was the number of rules with positive RHS, and *Negative score* was the number of rules with negative RHS. Then we divided *Positive score* by *Negative score* and then got a ratio we called *Positivity* . We only considered elementary clauses that occurred in at least 20 rules, and we classified those with positivity at least 4 as Positive, while those with positivity at most 1/4 as Negative.

To summarize our method: we searched for combinatorial biomarkers using a branch-and-bound algorithm for association rule mining; then made statistical analysis regarding elementary clauses.

3. Results

The program outputs over 200 rules from the CAMD database. Selected rules, with decreasing lift (i.e., "goodness") order, are listed in Table 4, (the whole set of rules are presented as Table S1 in the on-line supporting material).

Observe that, on the LHS, all clauses concerning biomarkers state that something is "too high" or "too low". That's because we thought that the best indicators can be values out of range. Normal values could probably indicate good health and thus normal cognition.

The first rule in Table 4 was that of the best lift: It can be interpreted in the following way: It is likely that if aspartate aminotransferase (AST) level is elevated, and subject is female, then her total MMSE score will be less than 15. Note that for all rules of ours do not necessarily mean a causal relation between the LHS and RHS, as both the LHS and RHS can be consequences of an unknown process in the background.

The second rule in Table 4 states that "if aspartate aminotransferase (AST) level is elevated, and subject is more than 65 years old, then her/his total MMSE score will be less than 15". The third rule states that "if serum sodium is elevated, and subject is more than 65 years old, then her/his total MMSE score will be less than 15".

From these rules we can conclude that elevated AST or sodium combined with relatively old age might be a good indicator (or even the cause) of mental decline.

Elementary clauses with positive effect on normal cognition are listed in Table 4.

Elementary clauses with negative effect on normal cognition (ordered by negativity increasing) are listed Table 5.

4. Discussion

Among the 238 rules identified, 51 rules had lift values exceeding 2.00. Those rules exceeding even the 3.00 lift value had one thing in common: the LHS contained the premise lb_ast=h These rules suggest that having higher levels of serum aspartyl aminotransferase (AST) may predispose to an impaired mental status characterized by a mini mental state examination score (MMSE) less than 15 points. Serum AST and alanyne aminotransferase (ALT) levels derive from the liver and their values may be elevated in a number of cases of liver injury or damage spreading from acute or chronic viral infections to alcohol induced or non-alcoholic steatohepatitis. It is interesting to note that elevated serum levels of AST, more than of ALT were associated with impaired mental status. Although mild elevations in serum levels of AST and ALT are nonspecific to the etiology of liver injury, certain alteration-patterns in these parameters may reflect the nature of the hepatic disease. For instance, the value of the AST/ALT ratio, also known as the De Ritis ratio is approximately 0.8 in normal subjects, however a ratio exceeding 2.00 is suggestive to alcoholic hepatitis. Therefore we scanned the subjects with high AST values for higher than 2 AST/ALT ratio: we have found only 10 subjects satisfying both conditions. Consequently, we may assume, that high serum AST in the study subjects are not typically accompanied with high De Ritis ratio, that may suggest alcoholic hepatitis.

The association of impaired liver function with mental decline can be illuminated by two perspectives. On one hand, impaired liver function might be insufficient to prevent the brain from the effects of certain neurotoxins e.g., ammonia. This happens in the case of hepatic encephalopathy (HE), when severe liver damage resulting in acute liver insufficiency cannot detoxificate ammonia and other neurotoxins. On the other hand, the association of elevated AST/ALT ratio with impaired mental status proposes that another

obscure element (e.g., chronic alcohol consumption) might be the factor responsible for both cognitive and metabolic damages. Our results raises the possibility of a pathogenetic linkage between liver function and mental status in patients with AD. Such linkage has also been proposed by other studies [20, 21]. One study concludes that peripheral reduction of β-amiloid is sufficient to reduce brain β-amiloid and proposes that β-amiloids, which are of major pathogenic importance in AD may originate from the liver [20]. Another research found that deficient liver production of a neuroprotective fatty acid, docosahexaenoic acid correlates with impaired cognitive status in AD patients [21].

Another identified rule states that among patients older than 65 years of age, higher levels of serum sodium concentrations increases the possibility with a 2.90 lift to achieve less than 15 points in MMSE. Net water loss is responsible for the majority of cases of hypernatremia [22]. A recent publication examining the causes and comorbidities in patients older than 65 years has found that the most common cause of community-acquired hypernatremia is dehydration due to reduced oral intake [23]. More interestingly, they found that the most common comorbidity in this patient group was AD, present in 31.4% of patients with hypernatremia [23]. Hydration status has a significant impact on the volumes of grey and white matter of the brain and on the quantity of the cerebrospinal fluid as a hallmark of ventricular enlargement [24]. The pattern of shrinkage in white matter volume and increase of the ventricular system due to dehydration is consistent with the structural brain changes observed during the progression of AD [24]. In another study, patients with AD underwent bioelectrical impedance vector analysis to assess the body cell mass and hydration status related to AD [25]. Results demonstrated a tendency towards dehydration in patients with AD [25]. Although the association of dehydration and AD is supported by these publications, the specific pathogenic nature of this association remains obscure [23, 24, 25].

More interestingly, we have found that some factors, considered to be risk factors in heart disease, may imply good cognitive status: we have found data of the positive effects of high serum cholesterol levels and high blood pressure.

We also need to mention the positive effects of high levels of B12 vitamin in blood serum.

It is not surprising that young age, high calcium, low chloride, low sodium have a positive effect on cognition. Male sex was probably classified as pos-

itive because Alzheimer's disease seems to be more frequent among women. Low pulse generally indicates good health, now it also seems to positively affect cognition. On the other hand, low temperature or high respiratory rate seem to contribute to mental decline (or at least indicate its presence). A more interesting part is that excess level of certain liver enzymes (AST, ALT and ALP) have a negative effect on cognition. It is possible that these high levels may be caused by some drugs treating dementia, especially Tacrine, but we should also consider the possibility that dementia and liver function might be related in some so far unknown way.

Another interesting result is that high Mean Corpuscular Hemoglobin and low Mean Corpuscular Hemoglobin Concentration seem to indicate normal cognition. This together means too large red blood cells with a lot of hemoglobin contained per cell. High eosinophil concentration or low lymphocite concentration also seem to be related to mental decline. High white blood count also seems to have a negative effect, probably because it indicates some infection, and it is not surprising that general bad health is in connection with cognitive impairment. High serum glucose also seems to be a negative factor. It has been suggested that diabetes contributes to mental decline.

References

[1] F. Bermejo-Pareja, J. Benito-Leon, S. Vega, M. J. Medrano, G. C. Roman, and Neurological Disorders in Central Spain (NEDICES) Study Group, "Incidence and subtypes of dementia in three elderly populations of central spain.", *J Neurol Sci* **264**(1-2), pp. 63–72 (2008).

[2] Antonio Di Carlo, Marzia Baldereschi, Luigi Amaducci, Vito Lepore, Laura Bracco, Stefania Maggi, Salvatore Bonaiuto, Egle Perissinotto, Guglielmo Scarlato, Gino Farchi, Domenico Inzitari, and I. L. S. A. Working Group, "Incidence of dementia, Alzheimer's disease, and vascular dementia in italy. the ilsa study.", *J Am Geriatr Soc* **50**(1), pp. 41–48 (2002).

[3] Marc Wortmann, "Dementia: a global health priority - highlights from an ADI and World Health Organization report.", *Alzheimers Res Ther* **4**(5), pp. 40 (2012).

[4] Martin Prince and Jim Jackson, "World Alzheimer Report 2009", Technical report Alzheimer's Disease International (2009).

[5] Enrico Mossello, Elena Ballini, Anna Maria Mello, Francesca Tarantini, David Simoni, Samuele Baldasseroni, and Niccolo Marchionni, "Biomarkers of Alzheimer's disease: from central nervous system to periphery?", *Int J Alzheimers Dis* **2011**, pp. 342980 (2010).

[6] Clifford R Jack, Val J Lowe, Stephen D Weigand, Heather J Wiste, Matthew L Senjem, David S Knopman, Maria M Shiung, Jeffrey L Gunter, Bradley F Boeve, Bradley J Kemp, Michael Weiner, Ronald C Petersen, and Alzheimer's Disease Neuroimaging Initiative, "Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease.", *Brain* **132**(Pt 5), pp. 1355–1365 (2009).

[7] Bruno Dubois, Howard H Feldman, Claudia Jacova, Steven T Dekosky, Pascale Barberger-Gateau, Jeffrey Cummings, Andre Delacourte, Douglas Galasko, Serge Gauthier, Gregory Jicha, Kenichi Meguro, John O'brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Salloway, Yaakov Stern, Pieter J Visser, and Philip Scheltens, "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria.", *Lancet Neurol* **6**(8), pp. 734–746 (2007).

[8] Sid E O'Bryant, Guanghua Xiao, Robert Barber, Joan Reisch, Rachelle Doody, Thomas Fairchild, Perrie Adams, Steven Waring, Ramon Diaz-Arrastia, and Texas Alzheimer's Research Consortium, "A serum protein-based algorithm for the detection of Alzheimer disease.", *Arch Neurol* **67**(9), pp. 1077–1081 (2010).

[9] Sid E O'Bryant, Guanghua Xiao, Robert Barber, Joan Reisch, James Hall, C. Munro Cullum, Rachelle Doody, Thomas Fairchild, Perrie Adams, Kirk Wilhelmsen, Ramon Diaz-Arrastia, Texas Alzheimer's Research, and Care Consortium, "A blood-based algorithm for the detection of Alzheimer's disease.", *Dement Geriatr Cogn Disord* **32**(1), pp. 55–62 (2011).

[10] K. Romero, M. de Mars, D. Frank, M. Anthony, J. Neville, L. Kirby, K. Smith, and R. L. Woosley, "The coalition against major diseases: developing tools for an integrated drug development process for alzheimer's

and parkinson's diseases.", *Clin Pharmacol Ther* **86**(4), pp. 365–367 (2009).

[11] Klaus Romero, Brian Corrigan, Christoffer W Tornoe, Jogarao V Gobburu, Meindert Danhof, William R Gillespie, Marc R Gastonguay, Bernd Meibohm, and Hartmut Derendorf, "Pharmacometrics as a discipline is entering the "industrialization" phase: standards, automation, knowledge sharing, and training are critical for future success.", *J Clin Pharmacol* **50**(9 Suppl), pp. 9S–19S (2010).

[12] James A Rogers, Daniel Polhamus, William R Gillespie, Kaori Ito, Klaus Romero, Ruolun Qiu, Diane Stephenson, Marc R Gastonguay, and Brian Corrigan, "Combining patient-level and summary-level data for alzheimer's disease modeling and simulation: a beta regression meta-analysis.", *J Pharmacokinet Pharmacodyn* **39**(5), pp. 479–498 (2012).

[13] David J. Hand, Heikki Mannila, and Padhraic Smyth, *Principles of Data Mining*, MIT Press (2001).

[14] Wei Wu, Wen Chun Juan, Cynthia R M Y Liang, Khay Guan Yeoh, Jimmy So, and Maxey C M Chung, "S100A9, GIF and AAT as potential combinatorial biomarkers in gastric cancer diagnosis and prognosis.", *Proteomics Clin Appl* **6**(3-4), pp. 152–162 (2012).

[15] Gabor Ivan, Zoltan Szabadka, and Vince Grolmusz, "Being a binding site: Characterizing residue composition of binding sites on proteins.", *Bioinformation* **2**(5), pp. 216–221 (2007).

[16] Christine Galustian and Angus G Dalgleish, "The power of the web in cancer drug discovery and clinical trial design: research without a laboratory?", *Cancer Inform* **9**, pp. 31–35 (2010).

[17] Richard J Epstein, "Unblocking blockbusters: using boolean text-mining to optimise clinical trial design and timeline for novel anticancer drugs.", *Cancer Inform* **7**, pp. 231–238 (2009).

[18] R. Chaves, J. M. Gorriz, J. Ramirez, I. A. Illan, D. Salas-Gonzalez, and M. Gomez-Rio, "Efficient mining of association rules for the early diagnosis of Alzheimer's disease.", *Phys Med Biol* **56**(18), pp. 6047–6063 (2011).

[19] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, "Mining association rules between sets of items in large databases", In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993* , pp. 207–216. ACM Press (1993).

[20] J Gregor Sutcliffe, Peter B. Hedlund, Elizabeth A. Thomas, Floyd E. Bloom, and Brian S. Hilbush, "Peripheral reduction of beta-amyloid is sufficient to reduce brain beta-amyloid: implications for Alzheimer's disease.", *J Neurosci Res* **89**(6), pp. 808–814 (2011).

[21] Giuseppe Astarita, Kwang-Mook Jung, Nicole C. Berchtold, Vinh Q. Nguyen, Daniel L. Gillen, Elizabeth Head, Carl W. Cotman, and Daniele Piomelli, "Deficient liver biosynthesis of docosahexaenoic acid correlates with cognitive impairment in Alzheimer's disease.", *PLoS One* **5**(9), pp. e12538 (2010).

[22] H. J. Adrogue and N. E. Madias, "Hypernatremia.", *N Engl J Med* **342**(20), pp. 1493–1499 (2000).

[23] Kenan Turgutalp, Onur Ozhan, Ebru Gok Oguz, Arda Yilmaz, Mehmet Horoz, Ilter Helvaci, and Ahmet Kiykim, "Community-acquired hypernatremia in elderly and very elderly patients admitted to the hospital: clinical characteristics and outcomes.", *Med Sci Monit* **18**(12), pp. CR729–CR734 (2012).

[24] Daniel-Paolo Streitbuerger, Harald E. Moller, Marc Tittgemeyer, Margret Hund-Georgiadis, Matthias L. Schroeter, and Karsten Mueller, "Investigating structural brain changes of dehydration using voxel-based morphometry.", *PLoS One* **7**(8), pp. e44195 (2012).

[25] R. Buffa, R. M. Mereu, P. F. Putzu, G. Floris, and E. Marini, "Bioelectrical impedance vector analysis detects low body cell mass and dehydration in patients with Alzheimer's disease.", *J Nutr Health Aging* **14**(10), pp. 823–827 (2010).

Tables

Table 1: Basic statistics on the subjects of the CAMD data

Age distribution		**Gender distribution**		**MMSE distribution**	
A: up to 65 years	1093	Female	3315	A: severe cog. impairment	255
B: 66-75 years	2070	Male	2653	B: moderate cog. impairment	611
C: 76-85 years	2408			C: mild cog. impairment	3224
D: more than 85	397			D: normal cognition	1352

Table 2: Several association rules of the highest lift

```
sex=F & lb_ast=h ---> mm_total=AB
Universe: 1546, LHS support: 96, RHS support: 346, Support: 72
Confidence: 0.75, Lift: 3.35116, Leverage: 0.0326746, X^2 stat: 50.4035
3.35116

lb_ast=h & age=BCD ---> mm_total=AB
Universe: 1546, LHS support: 117, RHS support: 346, Support: 84
Confidence: 0.717949, Lift: 3.20794, Leverage: 0.0373965, X^2 stat: 40.3832
3.20794

lb_sodium=h & age=BCD ---> mm_total=AB
Universe: 2477, LHS support: 100, RHS support: 556, Support: 65
Confidence: 0.65, Lift: 2.89577, Leverage: 0.0171794, X^2 stat: 80.1384
2.89577

lb_vitb12=h & age=ABC ---> mm_total=D
Universe: 3115, LHS support: 115, RHS support: 919, Support: 79
Confidence: 0.686957, Lift: 2.32848, Leverage: 0.0144694, X^2 stat: 60.3522
2.32848

lb_mch=h & age=BCD ---> mm_total=D
Universe: 2604, LHS support: 93, RHS support: 792, Support: 65
Confidence: 0.698925, Lift: 2.29798, Leverage: 0.0140992, X^2 stat: 59.7889
2.29798

lb_mchc=l & lb_chol=h ---> mm_ori=E
Universe: 1302, LHS support: 84, RHS support: 542, Support: 74
Confidence: 0.880952, Lift: 2.11624, Leverage: 0.0299787, X^2 stat: 18.3813
2.11624

lb_wbc_blood=h & (lb_bili=h or bpdia=h) ---> mm_attcal=B
Universe: 2821, LHS support: 132, RHS support: 1063, Support: 70
Confidence: 0.530303, Lift: 1.40732, Leverage: 0.00718192, X^2 stat: 31.7198
1.40732
```

Table 3: Legends for Table 2

sex=F	Subject is female
lb_ast=h	Serum Aspartate Aminotransferase level is too high
age=BCD	Subject is more than 65 years old
lb_sodium=h	Serum sodium is too high
lb_vitb12=h	Serum B12 vitamin is too high
age=ABC	Subject is at most 85 years old
lb_mch=h	Mean Corpuscular Hemoglobin is too high
lb_mchc=l	Mean Corpuscular Hemoglobin Concentration is too low
lb_chol=h	Serum cholesterol is too high
lb_wbc_blood=h	White blood count is too high
lb_bili=h	Serum indirect bilirubin is too high
bpdia=h	Diastolic blood pressure is too high
mm_total=AB	MMSE total score is less than 15
mm_total=D	MMSE total score is at least 24
mm_ori=E	MMSE orientation score is at least 8
mm_attcal=B	MMSE attention and calculation score is at most 1

Table 4: Elementary clauses with positive effect on normal cognition (ordered by positivity decreasing)

lb_mchc=l	Mean Corpuscular Hemoglobin Concentration is too low
lb_mch=h	Mean Corpuscular Hemoglobin is too high
age=ABC	Subject is at most 85 years old
sex=M	Subject is male
pulse=l	Pulse is too low

Table 5: Elementary clauses with negative effect on normal cognition (ordered by negativity increasing)

age=BCD	Subject is more than 65 years old
lb_eos=h	Eosinophils (particle concentration) is too high
lb_lym=l	Lymphocites (particle concentration) is too low
lb_wbc_blood=h	White blood count is too high
lb_gluc=h	Serum glucose is too high
lb_alp=h	Serum alkaline phosphatase (ALP) is too high
lb_cl=h	Serum chloride is too high
lb_ca=l	Serum calcium is too low
age=D	Subject is more than 85 years old
age=CD	Subject is more than 75 years old
lb_sodium=h	Serum sodium is too high
temper=l	Temperature is too low
resp=h	Respiratory rate is too high
lb_alt=h	Serum alanine aminotransferase (ALT) is too high
sex=F	Subject is female
lb_ast=h	Serum aspartate aminotransferase (AST) is too high