Intelligent Data Analysis for the Diagnosis of Alcohol Dependence Syndrome

D Dinevski1, P Povalej2 and M Kravos1

1Faculty of Medicine, and 2Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

Alcohol dependence syndrome is hard to diagnose because none of the existing laboratory markers alone has sufficient specificity and sensitivity. This study investigated whether combinations of markers would improve the identification of patients with alcohol dependence syndrome. Intelligent data analysis was carried out using the decision tree induction method with training and test data from 244 healthy volunteers and 238 patients with alcohol dependence syndrome. The results showed that a combination of two or three laboratory markers can identify alcohol dependence syndrome with almost 85% accuracy. It must be noted that induced decision trees offer a qualitatively different diagnostic evaluation of laboratory findings that varies from common practice, because they set up their own new borders and criteria that are different to generally accepted or set reference values. Tests for all of the selected laboratory markers are widely available, inexpensive to perform and usually form part of a routine laboratory examination.

KEY WORDS: Alcohol dependence syndrome; Markers of alcohol dependence syndrome; Data analysis; Intelligent systems; Decision trees

Introduction
The harmful use of alcohol, and alcohol dependence syndrome, with their variety of symptoms, are both disorders that are similar to other mental disorders.1 There is an association between behavioural, cognitive and physiological phenomena that result from recurrent and long-lasting alcohol abuse.2 Alcohol dependence syndrome might also lead to various complications such as withdrawal states and psychoses.1 Approximately 23 million European Union (EU) citizens have alcohol dependence syndrome.3

Early detection of alcohol dependence syndrome is vitally important because one-third of patients who have mental disorders also have health problems caused by excessive alcohol consumption.4 Moreover, alcohol dependence syndrome can lead patients with this disease to harm other people, most commonly family and friends, but also strangers. A total of 25% of road fatalities in the EU are alcohol related, which is the equivalent of 10 000 fatalities per year.1 Since alcohol dependence is one of the most widely spread addictions in the EU, the modernization of medical care in Europe with creation of a European-led market in healthcare5 could be one of the key factors
Diagnosing alcohol dependence by intelligent data analysis

Modern concepts of disorders caused by excessive alcohol consumption are based on categorical, rule-governed systems of diagnosis, found in the International Statistical Classification of Diseases (ICD-10)\(^6\) and the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV).\(^7\) The diagnosis of harmful alcohol consumption and alcohol dependence syndrome can be carried out using detailed history-taking, questionnaires and biochemical tests.\(^1\)

Ethanol metabolism is rather rapid when compared with that of many other drugs and it is essentially complete.\(^8\) The products of ethanol metabolism are carbon dioxide and water, which are indistinguishable from the products of the metabolism of foodstuffs and body energy stores. It is necessary, therefore, to use other markers of excessive alcohol consumption. No reliable biochemical markers for alcohol consumption and alcohol dependence syndrome have a sensitivity and specificity high enough to be relied upon to identify patients at risk of excessive consumption. The application of biochemical tests has its limitations, which should be understood and acknowledged. From the currently available laboratory tests, carbohydrate-deficient transferrin (CDT) offers the best combination of specificity and sensitivity.\(^9,10\) Other biochemical tests that are less reliable but are widely used in clinical practice include $\gamma$-glutamyltransferase (GGT), aspartate aminotransferase (AST), alanine transaminase (ALT), erythrocyte mean cell volume (MCV) and, rarely, glutamate dehydrogenase (GLDH). These markers are neither specific nor sensitive enough to detect patients with a chronic excessive alcohol intake.\(^11-17\) Thus, a combination of several markers might be more useful and this strategy is often used, although combinations are still not specific enough at providing biological proof of alcohol dependence syndrome.\(^16\)

To date, Stamm et al.\(^12\) has published the most systematic analysis of the combined use of several clinical chemical and haematological markers for the laboratory evaluation of alcohol dependence syndrome. The study considered five or more markers in combination and found specificities and sensitivities that ranged from 50 to 90%.\(^12\) The combinations of clinical chemical and haematological markers were, however, more effective at eliminating non-alcoholics than identifying patients with alcohol dependence syndrome.\(^12\) Moreover, Salaspuro\(^17\) found similar results in a meta-analysis of studies that compared CDT with either conventional or new biological markers of alcoholism, heavy drinking or alcohol use. Other authors have mainly focused on combinations of two laboratory markers.\(^18-20\)

Because of the lack of conventional methods for the successful diagnosis of alcohol dependence syndrome, intelligent data analysis methods were used to determine the most important laboratory markers or their combinations. An advanced multimethod approach was used to induce decision trees that can transform extracted knowledge into a humanly understandable form.\(^21\)

Patients and methods

**PATIENTS**

In-patients were recruited from the Psychiatric Hospital Ormož, Ormož, Slovenia, between September 2004 and July 2006. The study included only patients with alcohol dependence syndrome who met the following diagnostic criteria according to ICD-10:\(^6\) a strong desire to consume alcohol; difficulties in controlling alcohol use;
persistent alcohol consumption despite harmful consequences; a higher priority given to alcohol consumption than to other activities and obligations; increased tolerance; and sometimes a physical withdrawal state. Only in-patients were included in the study, because of the more reliable history provided by these patients and their relatives. The control group was selected from healthy patients and blood donors from general practices in Ormož and from the Department of Transfusion, General Hospital Ptuj, Slovenia, in the same region of Eastern Slovenia as the patients. Study participants were 18 – 65 years and of either sex. The exclusion criteria for both groups were: acute right heart failure; toxic circulatory failure; obstructive jaundice and severe respiratory insufficiency; acute viral hepatitis; severe acute intoxication; and hypovolaemic shock. Venous blood samples were taken using routine methods and were analysed immediately. One blood sample was taken from every patient on admission to hospital. Blood was taken from the healthy control group as a part of a regular examination in general practice or before blood donation.

The study protocol was approved by the National Medical Ethics Committee of the Republic of Slovenia, Ljubljana, Slovenia. All patients and healthy volunteers provided verbal informed consent.

INTELLIGENT DATA ANALYSIS
Intelligent data analysis has already been successfully used for knowledge extraction from many medical databases. Historically, different approaches for knowledge extraction have evolved, such as symbolic approaches and computational learning theory. Among them are many classical approaches including decision trees, decision rules, rough-sets, case-based reasoning, neural networks, support vector machines, different fuzzy methodologies, and ensemble methods, but they all have advantages and limitations. Evolutionary approaches are also a good alternative, because they are not inherently limited to local solutions. Taking into account the limitations of classical approaches, many researchers have focused their research on hybrid approaches, following the assumption that only the synergetic combination of single models can unleash their full power.

Selection of the appropriate method for data analysis heavily depends on the domain, the data and many other factors. Thus, for a given problem, different methods should be tested to increase the quality of the extracted knowledge. The representation of the extracted knowledge is, however, often one of the main criteria for choosing a knowledge extraction method. For example neural networks are known to produce highly accurate results but their drawback is that they do not explain how they have reached a particular conclusion. Thus, they are often used for clustering, sequencing and predicting patterns in clinical diagnosis, image and signal analysis, drug development, and so on. On the other hand, the representation of the extracted knowledge in the form of decision rules – or the more expanded form of decision rules, decision trees – is simple and easily understood by non-expert users. For that reason, the intelligent data analysis in the present study was limited to using decision tree-based methods for the knowledge extraction.

DECISION TREES
Decision trees have been successfully used for years in many decision-making applications. One of the main advantages of using decision trees, when
compared with other methods of machine learning, is that they provide a very simple and clear representation of the path taken to reach the acquired decision. Inducing a decision tree is a form of machine learning, where knowledge is extracted from a set of examples (objects) and presented in the two-dimensional form of a decision tree.30

A decision tree is induced on a training set, which consists of training objects. Every training object is completely described by a set of attributes (object properties) and a class (decision, outcome). Attributes can be numeric or discrete, but numeric attributes are not suitable for inducing a tree; therefore, they must be mapped into a discrete space. There are two types of nodes in a decision tree: internal and external nodes. Each internal node (non-terminal node) contains a test of a specific attribute value. External nodes (terminal nodes, decision nodes, leaves) are labelled with a class, which represents a decision. Nodes are connected with edges (links). Edges are labelled with different outcomes of a test performed on an attribute in a source node.

An induced decision tree is tested using a testing data set. The testing set consists of testing objects described with the same attributes as the training objects, except that the testing objects are not included in the training set.

Various methods can be used in the process of inducing a decision tree. The classical approach is to apply a heuristic function for the selection of the most important attribute. In the present study, in addition to the variety of heuristic functions, some advanced methods, like evolutionary algorithms and hybrid methods with neural networks and boosting, were used.

The accuracy of an induced decision tree is usually expressed with sensitivity and specificity. Specificity is defined as the number of healthy control subjects that were identified correctly as healthy out of the total number of healthy control subjects. Sensitivity is the number of correctly classified patients with alcohol dependence syndrome, divided by the total number of patients with alcohol dependence syndrome. The overall quality of an individual decision tree is described as total accuracy (equation 1) and average class accuracy on the test set (equation 2, below):

Equation 1:
\[ \text{Accuracy} = \frac{\text{No. of correctly classified objects}}{\text{No. of all objects}} \]

Equation 2:
\[ \text{Accuracy}_c = \frac{\text{No. of correctly classified objects in class } c}{\text{No. of all objects in class } c} \]

Average class accuracy \( \text{Average class accuracy}_c = \frac{\sum \text{Accuracy}_c}{\text{No. of classes}} \)

GENETICALLY INDUCED DECISION TREES

The disadvantages of classic decision tree induction, such as sensitivity to noise (missing or corrupted data),31 encouraged machine-learning researchers to try another method of machine learning that combines two methods: decision tree induction and genetic algorithms. This hybrid method merges the advantages of both methods and usually gives better results.

Genetic algorithms are based on the evolutionary ideas of natural selection and genetic processes of biological organisms.26,29 They are often capable of finding optimal solutions even in the most complex search spaces, or at least they offer significant benefits over other search and optimization techniques. The first phase of the genetic process is the generation of an initial population. Enough individuals have to be constructed to represent the whole
population. Every individual in this method is represented as a decision tree.

The second phase of the genetic process is the evolution of the population using three genetic operators. The first genetic operator is selection, where individuals are evaluated on the basis of fitness function and the best ones (parents) are chosen to create new individuals (children) with a second genetic operator known as crossover. In this way a new population is created. After the new individual is constructed by crossover, a genetic operator of mutation is applied with certain (low) probability. Mutation serves randomly to change individuals, with the intention of finding an optimal solution to the given problem in a faster and more reliable manner.

MUTIMETHOD APPROACH
The multimethod approach\(^2\) introduces the idea of a population of different intelligent systems, i.e. individuals that can produce multiple comparable good solutions, which are incrementally improved using the evolutionary approach. In order to enable knowledge sharing between different methods, support for the transformation between each individual method is provided. The initial population of intelligent systems is generated using different methods. In each generation different operations appropriate for individual knowledge representation are applied to improve the existing intelligent systems and also to create new intelligent systems. This enables incremental refinement of the extracted knowledge, with different views on a given problem. For example, using different induction methods, such as different purity, means that measures can be simply combined into a decision tree. As long as the knowledge representation is the same, a combination of different methods is not a big obstacle.

The main problem is how to combine methods that use different knowledge representations (for example neural networks and decision trees). In such cases, two alternatives are possible: (i) to convert one knowledge representation into another using different existing methods; or (ii) to combine both knowledge representations into a single intelligent system. The first alternative requires the implementation of knowledge conversion (for example conversion of a neural network into a decision tree). Such conversions are not perfect and some of the knowledge is normally lost, but conversions can produce a different aspect on a presented problem that can lead to better results. The second alternative requires some cut-points where knowledge representations can be merged. In a decision tree, internal nodes or decision leaves represent such cut-points, i.e. a condition can be replaced by another intelligent system (e.g. a support vector machine). Such trees are called hybrid decision trees. In contrast to conventional hybrids, the idea of the multimethod approach is to combine and apply different methods dynamically, in no predefined order, to solve a single problem or the decomposition of that problem (Fig. 1).

LABORATORY ANALYSES
Serum measurement of GGT, GLDH, AST and ALT were measured, and MCV was measured in whole blood. According to the recommendations of the International Federation of Clinical Chemistry, GGT, AST and ALT were ascertained with reference procedures at 37 °C.\(^3\) MCV was determined with the Abbott Cell-Dyn 610 (Abbott Laboratories, Abbott Park, IL, USA) and Melet Schloesing MS4 (Melet Schloesing Laboratoires, Osny, France) haematology analysers. The GLDH activity was measured using the Deutsche Gesellschaft für Klinische...
Chemie method and a Dialab GLDH kit (Dialab, Neudorf, Austria).

**STATISTICAL ANALYSES**

Statistical analyses were carried out using the SPSS® statistical package, version 12.0.1 (SPSS Inc., Chicago, IL, USA) for Windows®. Frequencies, arithmetic mean, median, variance, SD and SE of the arithmetic mean, highest and lowest values, specificity and sensitivity were calculated. The data were analysed using a Student’s t-test, the non-parametric Wilcoxon’s rank sum test and Mann–Whitney U-tests. A P-value of < 0.05 was considered to be statistically significant.

**Results**

The study comprised a control group of 244 healthy volunteers (199 men [81.6%]; 45 women [18.4%]) and 238 patients with alcohol dependence syndrome (199 men [83.6%]; 39 women [16.4%]). Among the healthy control subjects, there were 149 (61.1%) teetotallers. The mean ± SD age of the healthy control subjects was 44.9 ± 11.9 years, compared with 44.3 ± 8.8 years for the patients with alcohol dependence syndrome. There were no significant differences between the two groups regarding gender distribution and mean age.

The results of clinical chemistry and haematological analyses undertaken in the two groups, by gender, are shown in Table 1. The specificity and sensitivity of all of the biological markers are shown in Table 2. The statistical differences in clinical chemistry and haematological markers between the group of healthy control subjects and the group of patients with alcohol dependence syndrome are shown in Table 3; the non-parametric Mann–Whitney U-test of two independent samples showed that all measured clinical chemistry and haematological markers were statistically significantly higher in the group of patients with alcohol dependence (P < 0.001).

In order to perform the intelligent data analysis, the complete dataset was randomly divided into training and test sets at a ratio 2:1 and the results presented as specificity, sensitivity and total accuracy.
Diagnosing alcohol dependence by intelligent data analysis

The most interesting decision trees are presented in Figs 2 – 5. Decision tree (a) (Fig. 2), which was induced using the multimethod approach, accurately classified > 83% of the 161 test cases using two laboratory markers (GGT and MCV). The classification of patients with alcohol dependence syndrome in the test set (76 cases) (i.e. the sensitivity) was accurate in 80.26% and the classification of people without alcohol dependence syndrome was inaccurate for 12 cases out of 83 (specificity 85.88%).

The second decision tree (b) (Fig. 3) was slightly more accurate (84.47%) at diagnosing alcohol dependence syndrome than decision tree (a) by using the additional laboratory marker GLDH with GGT and MCV. The inclusion of the additional marker increased the sensitivity to 90.79%, but the specificity decreased to 78.82%.

The third decision tree (c) (Fig. 4) had an accuracy for diagnosing alcohol dependence syndrome of 82.61% using two laboratory markers (GGT and GLDH). This decision tree had a sensitivity of 89.47% and a specificity of 78.84%.

The fourth decision tree (d) (Fig. 5) had an accuracy for diagnosing alcohol dependence syndrome of 85.09% using four laboratory markers (MCV, AST, GGT and ALT). This decision tree had a sensitivity of 83.87% and a specificity of 86.76%.

Discussion

The present study demonstrated that the mean age of the patients with alcohol dependence syndrome was 44.3 years. This value was probably as high as this due to the relatively long period of at least 10 – 20 years that is necessary for the development of this disorder. It was also demonstrated that GLDH, AST, ALT, GGT activities and the MCV value of the patients with alcohol dependence syndrome were significantly higher than
Biochemical tests for the markers of alcohol dependence syndrome are routinely available, but none is 100% accurate. GLDH was the most specific laboratory marker identified in the present study, with a specificity of almost 90%. The highest sensitivity was achieved with MCV (87.2%), but its specificity was quite low (71.7%). This means that GGT and GLDH offered the best accuracy of the studied markers.

Among a huge number of induced decision trees, four of the most applicable decision trees that could be used to diagnose alcohol dependence syndrome were chosen because of their simplicity, easy examination and their moderate cost. We estimate that three diagnostic markers could be used to exclude eventual comorbid diseases or drug influences. Decision tree (b), which used the markers MCV, GGT and GLDH (Fig. 3), gave a model with almost 85% accuracy, an excellent sensitivity of > 90%, and a very high specificity of almost 80%. We feel confident about this model’s reliability and simple application, because it only requires three biochemical markers. This decision tree offers a qualitatively different approach to the diagnostic evaluation of laboratory findings that varies from common practice, because it sets up its own new borders and criteria that are different to generally accepted or set reference values. The MCV value is the leading step (branch) in the tree, which is very important. MCV rarely plays an important role in the diagnosis of alcohol dependence syndrome, as it is increased in many haematological diseases. This must be taken into account when considering the use of this decision tree for diagnosing alcohol dependence syndrome. Indeed, MCV had relatively low specificity (71.7%) for alcohol dependence syndrome. All three selected markers are widely available and inexpensive to measure. The high level of accuracy of this classification model suggests that it could be used to identify and diagnose patients with alcohol dependence syndrome in everyday practice, as long as the exclusion and differential diagnostic criteria used in the present study are taken into consideration.

Decision tree (a) included only two markers (GGT and MCV) and was less accurate (at just over 83%). The regular upper reference value in Slovenia for MCV (94 fl) was very near to the decision boundary (95.95 fl) found in decision tree (a), but some suggest 96 fl as the normal upper reference value, which should be more reliable.\textsuperscript{35} We believe that decision tree (a) could be more helpful as an additional screening test, rather than as a diagnostic test, for alcohol dependence syndrome.

Decision tree (c), with GGT and GLDH

<table>
<thead>
<tr>
<th>Marker</th>
<th>Specificity, %</th>
<th>Sensitivity, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCV</td>
<td>71.7</td>
<td>87.2</td>
</tr>
<tr>
<td>GGT</td>
<td>84.9</td>
<td>77.3</td>
</tr>
<tr>
<td>GLDH</td>
<td>89.8</td>
<td>65.5</td>
</tr>
<tr>
<td>AST</td>
<td>89.3</td>
<td>68.9</td>
</tr>
<tr>
<td>ALT</td>
<td>86.1</td>
<td>55.5</td>
</tr>
</tbody>
</table>

MCV, erythrocyte mean cell volume; GGT, $\gamma$-glutamyltransferase; GLDH, glutamate dehydrogenase; AST, aspartate aminotransferase; ALT, alanine aminotransferase.
included is, with its own boundary values close to generally accepted reference values, also close to diagnostic decision-making in regular clinical work. It could be very helpful in practices where GLDH is used as a regular laboratory marker.

Decision tree (d) was more sophisticated as it applied four common laboratory markers. Normally, AST would not have an important role to play in the diagnosis of alcohol dependence syndrome. The particularity of this decision tree was the exclusion criteria for non-alcohol dependents given to increased AST activity, and the necessity for further decisions at a low AST activity at step two. This decision strongly deviates from common clinical practice and it could provide a significant difference in the diagnosis of alcohol dependence syndrome, particularly if the fact that none of the individual markers or marker combinations reached an adequate level of specificity and sensitivity is taken into consideration. In clinical practice, however, biological markers often differ from clinical findings. The particularity of this tree was the boundary GGT value that extended deep into the otherwise normal reference activity.

The specificities, sensitivities and accuracies of all four chosen decision trees were within ranges comparable with other studies, based on the well-accepted biological reference ranges. \[11,35\] It was ascertained in the present study that the computer-based multimethod approach can produce multiple applicable solutions that are in some ways completely different to the biological ones.

In conclusion, the present study showed that intelligent data analysis can be very useful in diagnosing alcohol dependence syndrome and discovering new/better decision boundaries and biochemical marker combinations. The extracted knowledge confirmed several of the medical

### TABLE 3:

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>MCV</th>
<th>GGT</th>
<th>GLDH</th>
<th>AST</th>
<th>ALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mann-Whitney U-test</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Wilcoxon's rank sum test</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Asymptotic significance (two-tailed)</td>
<td>P &lt; 0.001</td>
<td>P &lt; 0.001</td>
<td>P &lt; 0.001</td>
<td>P &lt; 0.001</td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>MCV, erythrocyte mean cell volume; GGT, γ-glutamyltransferase; GLDH, glutamate dehydrogenase; AST, aspartate aminotransferase; ALT, alanine aminotransferase.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
facts that have been used to diagnose alcohol dependence syndrome. More importantly, some interesting new patterns were observed, which could help physicians and influence their decisions by providing exclusion criteria for their consideration.

Conflicts of interest
The authors had no conflicts of interest to declare in relation to this article.
FIGURE 4: Decision tree (c) for diagnosing alcohol dependence syndrome using γ-glutamyltransferase (GGT) and glutamate dehydrogenase (GLDH) (data in parentheses indicates the distribution of the training/test sets):

- **GGT** (322/161)
  - GGT ≤ 0.85 µkat/l
  - 0.85 µkat/l < GGT ≥ 1.49 µkat/l
  - GGT > 1.49 µkat/l
  - NO (165/91)
  - GLDH (34/18)
  - YES (123/52)
  - GLDH ≤ 125 nkat/l
  - GLDH > 125 nkat/l
  - NO (18/13)
  - YES (16/5)

FIGURE 5: Decision tree (d) for diagnosing alcohol dependence syndrome using erythrocyte mean cell volume (MCV), aspartate aminotransferase (AST), γ-glutamyltransferase (GGT) and alanine aminotransferase (ALT) (data in parentheses indicate the distribution of the training/test sets):

- **MCV** (322/161)
  - MCV ≤ 82.75 fl
  - MCV ≤ 94.5 fl
  - 94.5 fl ≤ MCV < 106.25 fl
  - MCV ≥ 106.25 fl
  - NO (2/1)
  - AST (119/50)
  - AST (147/89)
  - YES (3/6)
  - AST ≤ 0.83 µkat/l
  - AST > 0.83 µkat/l
  - AST ≤ 0.57 µkat/l
  - AST > 0.57 µkat/l
  - GGT (67/30)
  - YES (79/59)
  - GGT ≤ 0.38 µkat/l
  - GGT > 0.38 µkat/l
  - ALT (43/27)
  - ALT ≤ 0.71 µkat/l
  - ALT > 0.71 µkat/l
  - YES (38/25)
  - NO (5/2)
Diagnosing alcohol dependence by intelligent data analysis

• Received for publication 17 January 2011 • Accepted subject to revision 4 February 2011
• Revised accepted 2 April 2011

Copyright © 2011 Field House Publishing LLP

References


17 Salaspuro M: Carbohydrate-deficient transferrin as compared to other markers of alcoholism: a systematic review. Alcohol 1999; 19: 261 – 271.


Author’s address for correspondence
Professor Dejan Dinevski
Faculty of Medicine, University of Maribor, Slomškov trg 15, 2000 Maribor, Slovenia.
E-mail: dejan.dinevski@uni-mb.si