

Original Article

Data Mining Analysis of Demographic and Clinical Factors in Turkish Amyotrophic Lateral Sclerosis Patients

Nesrin Çelik Gülay, Hilmi Uysal¹, Pervin Aliyeva¹, Uğur Bilge

Department of Biostatistics
and Medical Informatics,
Akdeniz University,
¹Neurology Department,
Faculty of Medicine, Akdeniz
University, Antalya, Turkey

Submitted: 21-May-2020

Revised : 05-Sep-2020

Accepted : 01-Oct-2020

Published : 15-Jun-2021

ABSTRACT

Introduction: Amyotrophic lateral sclerosis (ALS) is a motor neuron disease that affects nerve cells in the brain and spinal cord, controlling voluntary muscle movement. Data mining is a discipline that provides meaningful conclusions from databases or implicit data. In this study, we examine the relationship between the clinical and demographic characteristics of ALS patients and a control group, using data mining techniques. **Methods:** In the study, data belonging to 235 patients diagnosed with ALS and a control group of 117 people consisting of relatives of ALS patients were used. The dataset contains 121 features that include clinical and demographic information for each patient. The patient group and the control group were examined together and separately to examine the relationship between the features. In the study the data mining methods of classification and clustering were used on R and WEKA software packages. **Results:** There were no significant differences between ALS patients and the control group in terms of environmental factors such as location, gender, smoking, exercise status, and clinical factors such as genetics, ALS involvement, course of the disease, disease in the family. The results also showed that there was no relationship between demographic and clinical features such as gender, occupation, age group, and concomitant disease between groups or within groups.

KEYWORDS: Amyotrophic lateral sclerosis, data mining, R package program, WEKA

INTRODUCTION

Amyotrophic lateral sclerosis (ALS) was described by the French neurologist Jean-Martin Charcot in 1874 as “a fatal neurodegenerative disease characterized by the selective loss of motor neurons in the cortex, brainstem and spinal cord.” ALS is a rare neurological disease in the brain and spinal cord that affects nerve cells (neurons), controlling voluntary muscle movement. Due to the large number of genetic and environmental factors that contribute to the pathogenesis of the disease, and because it involves complex interactions between genetic and/or environmental factors, it is one of the family of “complex diseases.”^[1-3]

The cases of ALS incidence in the world, which are stated as 1.2–4 per 100,000 population, have been reported as over 6 per 100,000 population as a result of studies conducted in Turkey.^[4-10] In approximately 10%

of ALS patients, the origin of the disease is genetic and half of these patients show familial inheritance. Although ALS can occur at any time in adulthood, it usually affects people in their mid-fifties and there are five known different phenotypes ALS, progressive muscular atrophy, primary lateral sclerosis, PBS and ALS-Frontotemporal dementia.^[6,11]

ALS affects people of all races and ethnic backgrounds. While the etiology of ALS has not been determined yet, studies show some potential risk factors for ALS. These are demographic factors such as age, gender, race and ethnicity; environmental factors such as

Address for correspondence: Dr. Uğur Bilge,
Department of Biostatistics and Medical Informatics,
Akdeniz University, Antalya 07000, Turkey.
E-mail: ubilge@akdeniz.edu.tr

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Gülay NÇ, Uysal H, Aliyeva P, Bilge U. Data mining analysis of demographic and clinical factors in Turkish amyotrophic lateral sclerosis patients. *Neurol Sci Neurophysiol* 2021;38:111-9.

Access this article online

Quick Response Code:



Website: www.nsnjournal.org

DOI: 10.4103/NSN.NSN_69_20

smoking, physical activity, place of residence, etc., and genetic and epigenetic factors.^[12-17] Studies have been conducted in areas such as establishing assistive systems for the diagnosis of the disease for ALS disease using information technologies and predicting the symptoms of the disease and the course of the disease.^[18-23]

In this study, the aim was to investigate the relationship between ALS diagnosed cases and the control group in terms of clinical, demographic and genetic factors using data mining methods.

METHODS

The sample of the study includes the control group consisting of healthy relatives of patients who were diagnosed with ALS according to the revised El Escorial criteria and who did not have a similar neurodegenerative disease (Parkinson's disease, Alzheimer's disease, etc.) after applying to the Neurology Department, Faculty of Medicine Akdeniz University.

The data of 235 patients and 117 healthy individuals who applied to Akdeniz University Hospital between 2016 and 2018 were studied, with the patients' informed consent given after obtaining the approval of Akdeniz University Faculty of Medicine Ethics committee. Data from individuals included in the study were collected prospectively using a standard questionnaire developed by the OnWebDuals consortium.

More than 500 questions were asked to the patient and control groups, including main questions and sub-questions. In these questions, demographic information (age, gender, place of residence characteristics, exercise, diet, occupation, etc.), family and environment related questions, disease related information, existence of other diseases, information about where in the body the disease started and how it spread, ALS type, as well as laboratory and imaging results, genetic mutation sites were registered. The answers given to the questions in the questionnaire were saved into an Excel spreadsheet file using check rules and filters. When the data was processed it was determined that there were many areas with incorrect or incomplete answers, and that patient and control groups could not answer all the questions. As a result of this, the questions were reviewed again and 121 questions or attributes were selected as all patients had answered and were suitable for analysis.

In order to analyze the data, Clustering and Decision Tree methods were used from R and WEKA data mining packages. K-Means Clustering and Hierarchical Clustering methods aimed for a maximum similarity within each cluster at the end of the partitioning process,

enabling data with similar characteristics to be grouped into same clusters. The Decision Tree method was used to help visualize the relationships between data by representing relationships with branches and leaves. We used both R and WEKA programs, which are both open source and contain many algorithms for data mining, to show that the results obtained from different software platforms are compatible. The Clustering and Decision Tree methods were preferred in order to facilitate the determination of the relationships between the features that could not previously be predicted with traditional approaches, as the ALS dataset is a high dimensional dataset with many attributes. We used multiple techniques to ensure that the separation between patient groups is verified with different approaches with maximum accuracy.

The data obtained have been processed according to CRISP-DM, one of the data mining methods, and made suitable for analysis.^[24] With the method used in 156 features, the method was reduced to 121 features for patients, 29 features for healthy people and made ready for analysis.

Although data mining, which enables the management of data and reveals meaningful information from data, is used as a common term today, it first entered the literature in the 1980s. Since then it has progressed rapidly under the headings of data mining, Statistical Science, Artificial Intelligence and Machine Learning. Choosing the right algorithm is an important parameter for successful data mining studies. In the selection of the algorithm, the type or structure of the data is of great importance. In many studies, data mining models and algorithms are discussed under different headings. The three most commonly used methods are Classification, Clustering, and Association Rules.^[25-29]

Classification and Clustering Methods were used in this study. In selected methods, the aim is to group, identify similarities in the dataset and classify the data into predefined or undefined classes. In order to use these methods, K-Means Clustering, Hierarchical Clustering and Decision Tree methods have been applied.

Statistical analysis

The dataset used in this study consisted of 235 patients each with 121 characteristics and 117 healthy individuals each with 29 characteristics. The difference between the percentages of categorical variables was analyzed by Pearson's Chi-Square Test. If more than 20% of the expected frequencies were <5, Fisher's Exact Test was used. A value of $P < 0.05$ was used to assess the significance for all statistical analyses. The

Statistical Package for the Social Sciences (SPSS) 16.0 (SPSS Inc.; Chicago, IL, USA) was used for statistical analysis, and the information obtained is given in the Results section.

Data mining analysis

Data Mining analysis was carried out on R 3.5.3 programming environment and WEKA 3.8.4 packages; both are an open source software, with many ready-made algorithms.^[30,31] The K-Means clustering algorithm groups data according to their similar characteristics, aiming for maximum distances between clusters as well as maximizing similarity within each cluster. Hierarchical Clustering and the Decision Tree methods are used to visualize by dividing data into smaller groups to form a tree structure based on similarity.^[26,32-35]

RESULTS

Although the age range of the dataset examined within the scope of the research is divided into productive and non-productive periods, the average age of the patients is 60 and 77.4% of the patients are in the non-productive period. However, among the patient group, the youngest patient was 24 years old and the oldest patient was 82 years old in terms of the age at which symptoms were seen. In addition, 62.1% of the patients (146) are male patients. When examined in terms of ALS phenotype, 99% of the patients (233 people) were recorded as ALS type. More detailed information about the descriptive characteristics of patients and healthy individuals is summarized in Table 1.

Data mining analysis results

K-means cluster analysis results

In the K-Means method used to find clusters or subgroups of observations in the dataset, elements in the same cluster are expected to be similar, but differ from the elements in different clusters. In this method, K represents the number of groups in the dataset. In order to determine K, the Silhouette Method was used to establish the optimum number of clusters. We ran tests on three different datasets. These are patients, controls, and patient + controls.

For the patient group, the optimum number of clusters was found as 5 clusters. For the control group, the optimum number of clusters was also 5 clusters. Finally, for the patient + control group the optimum number of clusters was found to be 2 clusters.

The first group to be examined according to the optimum cluster numbers is the ALS group and the results are given in Figure 1. The result for the patient group in Figure 1 visualizes the 5 different clusters and members of the cluster. Each cluster is shown

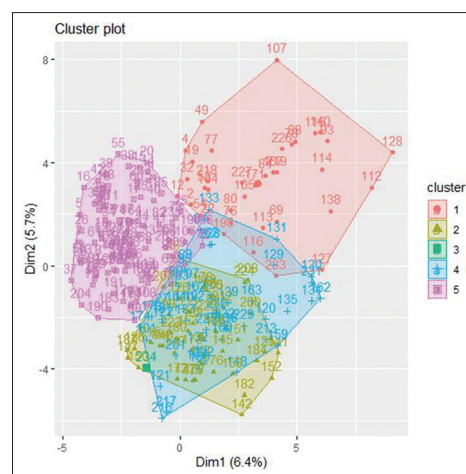


Figure 1: K-Means cluster analysis of the patient group (Visualization of the cluster analysis result. The patient group is divided into 5 groups according to the optimum cluster method with R program)

by a different color. Cluster results show that patient groups and group members are not segregated and 3 clusters are intertwined. For this reason, it was not possible to determine the attributes of the elements that show intra-cluster similarities and differences between clusters.

For the patient group, the same number of clusters was also studied in the WEKA program. In the WEKA program, patients were divided into 5 groups. The number and percentages of the separated patients were as follows: First group 39 people (17%), second group 65 people (28%), third group 52 people (22%), fourth group 28 people (12%), fifth group 51 people (22%).

The image formed in line with the optimum number of clusters for the control group is given in Figure 2.

Accordingly, it is seen that the elements of the two clusters diverge significantly, but the distances of the clusters are close to each other. When the cluster elements are examined, it can be said that the similarities of the second cluster are mostly due to the attributes “they do not have any disease, or in their family” and “place they live.”

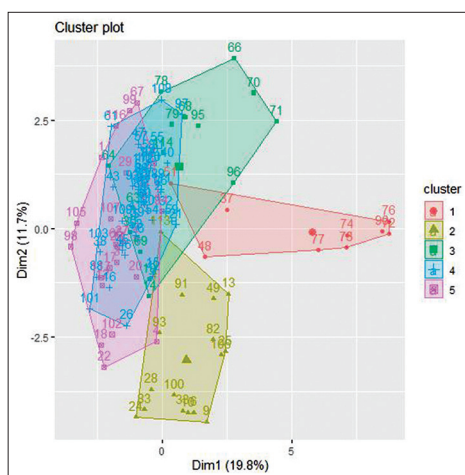
For the control group, it was also run in the WEKA program with the same number of clusters. The results obtained here give the number of people in each cluster; 29 (25%) for Cluster 1, 17 (15%) for Cluster 2, 26 (22%) for Cluster 3, 34 (29%) for Cluster 4, and 11 (9%) for Cluster 5.

The results obtained by examining the patient and control groups together are given in Figure 3.

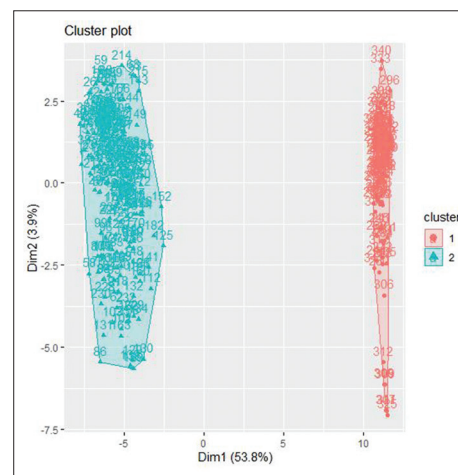
In the image, it is seen that the patient and control groups are distinctly separated, while the distances of the group members are mostly small. Group 1,

Table 1: Comparison of amyotrophic lateral sclerosis and control group in terms of defining characteristics

	ALS (n=235)	Control (n=117)	P
Sex			
Female	89 (37.9)	65 (55.6)	0.001 ^a
Male	146 (62.1)	52 (44.4)	
Age period			
Productive	53 (22.6)	52 (44.4)	<0.001 ^a
Nonreproductive	182 (77.4)	58 (49.6)	
Place of living features			
Village	86 (36.6)	19 (16.2)	<0.001 ^a
Country town	30 (12.8)	7 (6.0)	
Small town	35 (14.9)	17 (14.5)	
Middle town	30 (12.8)	15 (12.8)	
Large town	54 (23.0)	58 (49.6)	
Smoking			
Yes	49 (20.9)	38 (32.5)	0.001 ^a
No	159 (67.7)	77 (65.8)	
Diet (Mediterranean)			
Yes	4 (1.7)	2 (1.7)	0.052 ^b
No	231 (98.3)	112 (95.7)	
Occupation			
Armed forces occupations	12 (5.1)	-	-
Managers	5 (2.1)	1 (0.9)	
Retired/employee	76 (32.3)	8 (6.8)	
Professionals	19 (8.1)	15 (12.8)	
Technicians and associate professionals	7 (3.0)	3 (2.6)	
Clerical support workers	7 (3.0)	-	
Service and sales workers	20 (8.5)	4 (3.4)	
Skilled agricultural, forestry and fishery workers	29 (12.3)	3 (2.6)	
Craft and related trades workers	6 (2.6)	4 (3.4)	
Plant and machine operators, and assemblers	4 (1.7)	1 (0.9)	
Elementary occupation	22 (9.4)	20 (17.1)	
Exercise			
Yes	36 (15.3)	14 (12.0)	<0.001 ^a
No	98 (41.7)	20 (17.1)	

^aChi-square test, ^bFisher's exact test**Figure 2:** K-Means cluster analysis of the control group (Visualization of the cluster analysis result. The control group is divided into 5 groups according to the optimum cluster method with R program)

represented by red, represents the control group, and Group 2, represented by blue, represents the ALS

**Figure 3:** K-Means cluster analysis of the patient and control group (Visualization of the cluster analysis results. The patient + control group is divided into 2 groups according to the optimum cluster method with R program)

patient group. Here, the reason for the clear separation of the two groups is the absence of clinical signs of

the disease in the control group. Except for clinical features, features differing between groups could not be determined. In the cluster details, it was seen that the distance of some of the control group members in the 1st cluster was far from the others due to the differences in whether they are in the family, have other diseases, gender, the place of residence, the place of birth and the unproductive period, but we did not observe any difference between the patients in cluster 2.

For the patient and control groups, it was run in the WEKA program with the same number of clusters. The results obtained here were found to be divided into 2 clusters: 117 (33%) for Cluster 1 and 235 (67%) for Cluster 2.

Hierarchical cluster analysis results

Datasets for this method, which was performed according to similar properties of cluster elements, were examined in 3 groups as patient, control, and patient + control. The results for these groups are schematized with cluster plots and dendrograms.

Clustering and Dendrogram Graph results are given for the patient dataset in Figure 4.

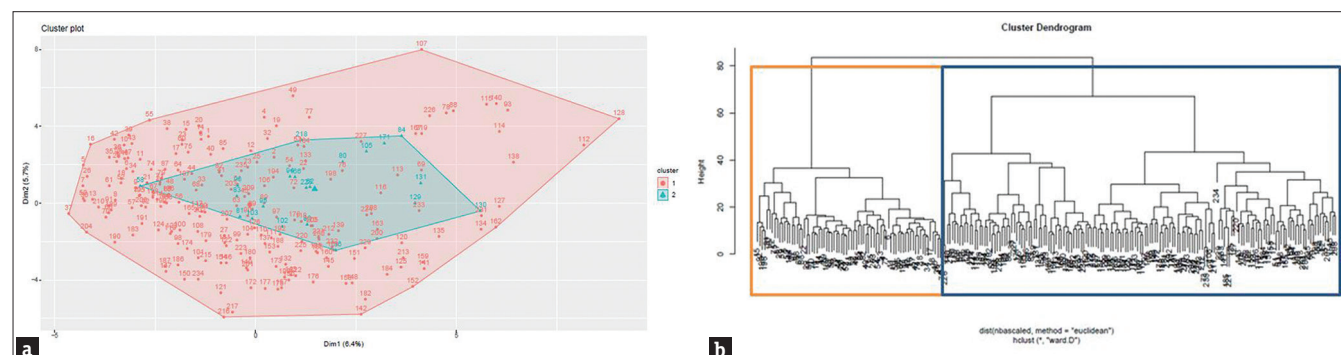


Figure 4: (a and b) Hierarchical clustering and dendrogram chart of the patient group (Clustering chart and Dendrogram chart obtained by using the R program as a result of dividing the patient group into two clusters)

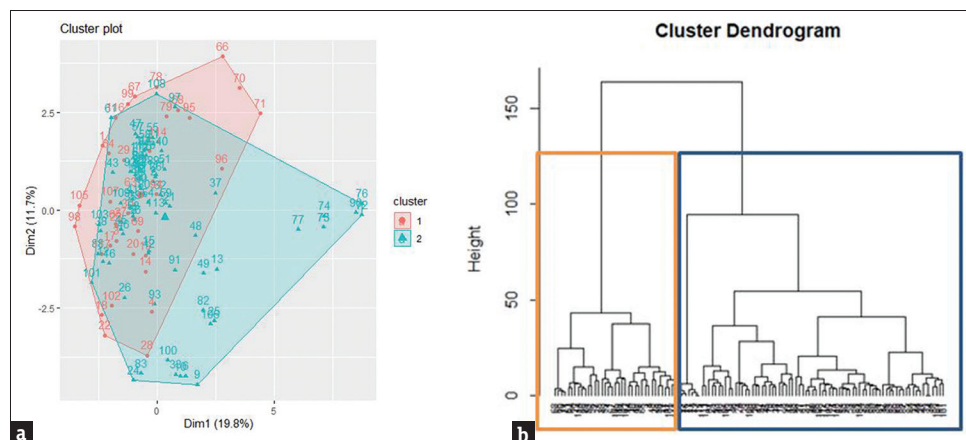


Figure 5: (a and b) Hierarchical clustering and dendrogram chart of the control group (Clustering chart and Dendrogram chart obtained by using the R program as a result of dividing the control group into two clusters)

According to the results obtained, the figure shown in red shows the 1st cluster and the figure shown in blue shows the 2nd cluster. In the cluster graph, it is seen that the patients are very similar in terms of features and nested clusters are formed that do not separate from each other.

For the patient group, the Hierarchical Clustering method was also run in the WEKA program. The results obtained here were divided into 2 clusters as 234 people (99%) in the first cluster and 1 person (1% as below) in the second cluster.

When the dataset consisting of control groups is analyzed by the Hierarchical Clustering method, the clustering graph and dendrogram graph are given in Figure 5.

According to the clustering graph, it is seen that 2 separate groups are formed. According to the results of the dendrogram, it is seen that the distance between the two group members is small, but the distance between the groups is high. This result indicates that the control group is similar in terms of its characteristics.

For the control group, the Hierarchical Clustering method was also run in the WEKA program. The results

obtained here were found to be divided into 2 clusters, 116 (99%) for Cluster 1 and 1 (1%) for Cluster 2.

According to the hierarchical clustering method, patient and control group data in R and WEKA were examined and the results were given in Figure 6.

As seen in the dendrogram in Figure 6, the patient and control groups are distinctly separated, while the members of each group are close to each other. Group 1 is formed by patients, and Group 2 has members of the control group. The clear separation of the two groups is due to the absence of clinical attributes of the disease in the control group. Apart from this, the two groups do not separate from each other in terms of other features.

Hierarchical Clustering method was also run in WEKA program for patient and control groups. The results obtained here were found to be divided into 2 clusters, namely 235 (61%) for Cluster 1 and 117 (39%) for Cluster 2.

Decision tree analysis results

For this method, which shows how the branches and leaves of the tree are formed in certain cases from the dataset, it was used for the patient + control group with the ALS patient group and the control group without the ALS patient. Other information about the disease status was added to the dataset and the patients with and without illness were coded. In both R and WEKA programs, no rule could be obtained with this method and a tree model consisting of nodes and leaves could not be formed.

As a result of the analyses, it has been determined that the similarities or differences between patients and control groups are mainly based on clinical findings rather than demographic information.

In the literature, ALS is more common in men than women. Similarly, in our study, the patient group

consisted of 62.1% male. In the total dataset, females were (154) 43.75% and the males were (198) 56.25%. There was no significant difference between genders in terms of the disease status ($P < 0.001$).

While smoking was reported as an important factor in the literature, smoking was not seen as a determining factor in our study, and no significant relationship was found between gender and disease status ($P < 0.001$).

It is seen in the literature there are discussions regarding physical activity and the hypothesis that physical activity is a risk factor for the development of ALS. According to results from our study, no significant difference was found in terms of daily life activity/regular exercise ($P < 0.001$).

In the study, dendrogram and clustering graphics were actively used to visualize the results and were seen as a factor that facilitates the data evaluation process.

In addition to these results, the Decision Tree method, which is one of the data mining classification models, was tried in the study, but a distinctive feature that enabled the formation of the tree did not occur.

DISCUSSION AND CONCLUSION

Although various risk factors are mentioned in studies on ALS, these factors are still being discussed as the disease is a complex disease. In this study, the relationships between the clinical and demographic characteristics of people with and without ALS disease were investigated using data mining methods.

It has been stated in many sources in the literature that the prevalence of ALS patients in men is higher than that of women. Therefore, the effect of gender differences on other characteristics was examined in detail in the study. According to the results of the study, the gender difference can not be explained by clinical, genetic and

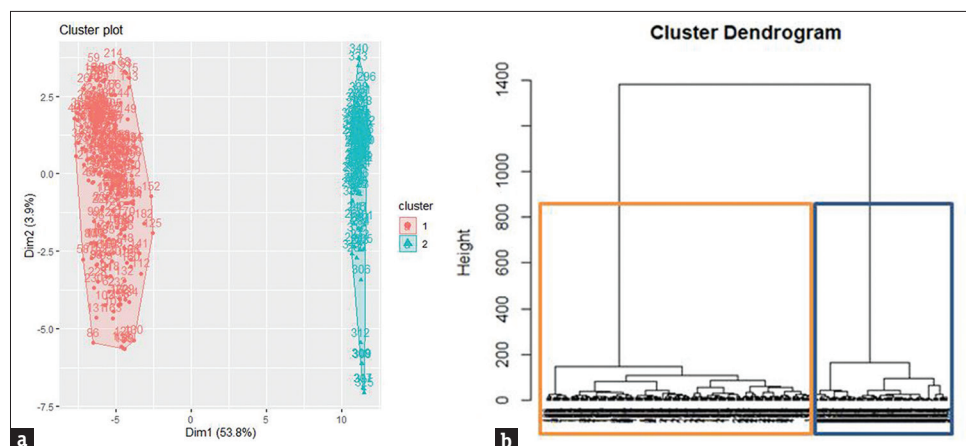


Figure 6: (a and b) Hierarchical clustering and dendrogram chart of the patient and control group (Clustering chart and Dendrogram chart obtained by using the R program as a result of dividing the patient and control group into two clusters)

other demographic data. As for the progression of the disease in terms of gender, no difference was found between men and women. In addition, it was determined that there was no difference in terms of environmental, demographic and other factors investigated. Since most of the patients included in the study had ALS in terms of phenotype the investigated environmental, demographic and other factors did not differ in the groups, in terms of this phenotype.

Smoking, which is stated as an important factor in the literature, was not found to be a determining factor in our study, when the patient and control groups were examined together.^[36,37] This could be because of the high number of patients with no data on smoking, or because some had already given up smoking, and this was not recorded in the dataset.

Although there is some evidence in the literature regarding the effect of physical activity intensity or regular exercise on the development of the disease, we found no significant difference between ALS and control groups in terms of physical activity.^[38-42] The reason for that is thought to be not having sufficient information about the exercise status of the patients.

The results obtained in the study emphasize the importance of clinical findings in diagnosing the disease. Clinical findings constituted the reason for the distinction between the patient and control groups. In addition, it was observed that minor differences between the patient group occurred in terms of clinical features, but in our study, this difference was not sufficiently explained by the factors specified in the literature such as genetic characteristics, gender differences, and environmental factors.^[43,44]

In the process of examining the relationship between patient and control groups in terms of clinical, genetic and demographic factors, R and WEKA programs and Classification, Clustering and Decision Tree methods were used. The results, indicating that there is no difference in terms of data relationships, were obtained within the scope of the study.^[26,32-35] Also, achieving similar results in both R and WEKA programs shows that the results can be verified. The results obtained show that other than these programs, different data mining tools or programs can be used and the obtained outputs can be analyzed comparatively.

The fact that ALS is a complex disease and may involve interactions of a large number of patient attributes, means that the main limitation of the study is the number of patients and the controls. We believe in the future there will be larger datasets compiled, and some of the data mining techniques we introduce here

will be used by researchers and physicians working in this field.

Despite the limitations of the study, some meaningful results were obtained with K-Means Clustering and Hierarchical Clustering methods, but no results were obtained with the Decision Tree method. The reason for this is thought to be due to the insufficient number of samples in the learning set.

Findings from the data mining analyses are an indication that as the data collection process continues and the number of cases increases, the clinical, demographic, genetic and epigenetic risk factors can give new clues about the progression of ALS. Therefore, it is thought that multicenter, national, and international studies including large datasets and control groups should be continued in order to support the results.

Acknowledgment

We would like to thank the OnWebDuals project group, which allowed ALS patient group data to be analyzed through interdisciplinary teams, and put great effort in the preparation of the questionnaire and the standardization of the dataset.

In addition, we would like to thank Research Assistant Cansu Aydın at Akdeniz University Faculty of Medicine Department of Neurology, and PhD Student Vildan Çiftçi at Akdeniz University Faculty of Medicine, Department of Medical Biology and Genetics who contributed to the preparation of the data within the scope of the study.

With the limited dataset compiled from the study, presentations were made at the International Participation XI Medical Informatics Congress, Korkut Yaltkaya XIV Workshop, and Clinical Neurophysiology Symposium.

The budget of the study for collecting the dataset used in the presented article was supported by Akdeniz University Scientific Research Projects Coordination Unit (BAP) with the code TTU-2017-2661.

We thank Philippa Price for proofreading the article.

The authors do not report any conflict of interest regarding this manuscript.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Eisen A, Krieger C. Amyotrophic Lateral Sclerosis: A Synthesis of Research and Clinical Practice. New York: Cambridge University Press; 1998.
2. National Institute of Neurological Disorders and Stroke. Office

- of Communications and Public Liaison Amyotrophic Lateral Sclerosis (ALS) Fact Sheet. Available from: <https://www.ninds.nih.gov>. [Last accessed on 2020 May 01].
3. Majoor-Krakauer D, Willem PJ, Hofman A. Genetic epidemiology of amyotrophic lateral sclerosis. *Clin Genet* 2002;63:83-101.
4. Benjaminsen E, Alstadhaug KB, Gulsvik M, Baloch FK, Odeh F. Amyotrophic lateral sclerosis in Nordland county, Norway, 2000-2015: Prevalence, incidence, and clinical features. *Amyotroph Lateral Scler Frontotemporal Degener* 2018;19:522-7.
5. Jun KY, Park J, Oh KW, Kim EM, Bae JS, Kim I, *et al.* Epidemiology of ALS in Korea using nationwide big data. *J Neurol Neurosurg Psychiatry* 2019;90:395-403.
6. Turgut N, Varol Saraçoglu G, Kat S, Balci K, GÜldiken B, Birgili O, *et al.* An epidemiologic investigation of amyotrophic lateral sclerosis in Thrace, Turkey, 2006-2010. *Amyotroph Lateral Scler Frontotemporal Degener* 2019;20:100-6.
7. Zhou S, Zhou Y, Qian S, Chang W, Wang L, Fan D. Amyotrophic lateral sclerosis in Beijing: Epidemiologic features and prognosis from 2010 to 2015. *Brain Behav* 2018;8:E01131.
8. Leighton DJ, Newton J, Stephenson LJ, Colville S, Davenport R, Gorrie G, *et al.* Changing epidemiology of motor neurone disease in Scotland. *J Neurol* 2019;266:817-25.
9. Rose L, McCim D, Lease D, Nonoyama M, Tandon A, Bai YQ, *et al.* Trends in incidence, prevalence, and mortality of neuromuscular disease in Ontario, Canada: A population-based retrospective cohort study (2003-2014). *PLoS One* 2019;14:E0210574.
10. Longinetti E, Fang F. Epidemiology of amyotrophic lateral sclerosis: An update of recent literature. *Curr Opin Neurol* 2019;32:771-6.
11. Uysal H, Taghiyeva P, Türkay M, Köse F, Aktekin M. Amyotrophic lateral sclerosis in Antalya, Turkey. A prospective study, 2016-2018. *Amyotroph Lateral Scler Frontotemporal Degener* 2020;1-7.
12. Hastings MH, Goedert M. Circadian clocks and neurodegenerative diseases: Time to aggregate? *Curr Opin Neurobiol* 2013;23:880-7.
13. Sutedja NA, Veldink JH, Fischer K, Kromhout H, Heederik D, Huisman MH, *et al.* Exposure to chemicals and metals and risk of amyotrophic lateral sclerosis: A systematic review. *Amyotroph Lateral Scler* 2009;10:302-9.
14. Hamidou B, Couratier P, Besançon C, Nicol M, Preux PM, Marin B. Epidemiological evidence that physical activity is not a risk factor for ALS. *Eur J Epidemiol* 2014;29:459-75.
15. Das K, Nag C, Ghosh M. Familial, environmental, and occupational risk factors in development of amyotrophic lateral scler. *North Am J Med Sci* 2012;4:350-5.
16. Wang MD, Little J, Gomes J, Cashman NR, Krewski D. Identification of risk factors associated with onset and progression of amyotrophic lateral sclerosis using systematic review and meta-analysis. *Neurotoxicology* 2017;61:101-30.
17. Zou ZY, Zhou ZR, Che CH, Liu CY, He RL, Huang HP. Genetic epidemiology of amyotrophic lateral sclerosis: A systematic review and meta-analysis. *J Neurol Neurosurg Psychiatry* 2017;88:540-9.
18. Rong P, Yunusova Y, Wang J, Green JR. Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach. *Behav Neurol* 2015;2015:183027.
19. Kafkafi N, Yekutieli D, Yarowsky P, Elmer GI. Data mining in a behavioral test detects early symptoms in a model of amyotrophic lateral sclerosis. *Behav Neurosci* 2008;122:777-87.
20. Pires S, Gromicho M, Pinto S, Carvalho M, Maderia SC. Predicting noninvasive ventilation in ALS patients using stratified disease progression groups. In: *IEEE Int Conf Data Mining Workshops*, pp. 74857 (2018).
21. Ning Z, Li L, Jin X. Classification of Neurodegenerative Diseases Based on CNN and LSTM. *Proceedings-9th International Conference on Information Technology in Medicine and Education*; 2018. p. 82-5.
22. Alaskar H, Hussain AJ. Data mining to support the discrimination of amyotrophic lateral sclerosis diseases based on gait analysis. *Lecture Notes Comp Sci* 2018;10956:7606.
23. Halberberg D, Lerner B. Temporal modeling of deterioration patterns and clustering for disease prediction of ALS patients. In: *IEEE Int Conf Mach Learning Appl*, pp.62-8 (2019).
24. Shearer C. The CRISP-DM model: The new blueprint for data mining. *J Data Warehousing* 2000;5:13-22.
25. Ramagari BM. Data mining techniques and applications. *Indian J Comp Sci and Eng* 2011;1:301-5.
26. Nikam SS. A comparative study of classification techniques in data mining algorithms. *Orient J Comp Sci Technol* 2015;1:13-9.
27. Nisbet R, Elder J, Miner G. *Handbook of Statistical Analysis and Data Mining Applications*. 1st ed. Canada: Academic Press Elsevier; 2009.
28. Gündoğdu ÖE. *Genetic Algorithms in Data Mining [dissertation]*. Kocaeli University; 2007.
29. Han J, Kamber M. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham: Morgan Kaufman Publishers; 2006.
30. Ihaka R, Gentleman RR. A language for data analysis and graphics. *J Computat Graphical Statist* 1996;5:299-314.
31. WEKA. WEKA The Workbench for Machine Learning; 2020. Available from: <https://www.cs.waikato.ac.nz/~ml/weka/>. [Last accessed on 2020 May 04].
32. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. *Pattern Recog* 2003;36:451-61.
33. Quinlan JR. Induction of decision tree. *Mach Learning* 1986;1:81-106.
34. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. *Bioinformatics* 2004;20:1453-4.
35. Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recog Lett* 2010;31:651-66.
36. Pleil JD, Stiegel MA, Madden MC, Sobus JR. Heat map visualization of complex environmental and biomarker measurements. *Chemosphere* 2011;84:716-23.
37. Alonso A, Logroscino G, Hernán MA. Smoking and the risk of amyotrophic lateral sclerosis: A systematic review and meta-analysis. *J Neurol Neurosurg Psychiatry* 2010;81:1249-52.
38. Oskarsson B, Horton DK, Mitsumoto H. Potential environmental factors in amyotrophic lateral sclerosis. *Neurol Clin* 2015;33:877-88.
39. Harwood CA, Westgate K, Gunstone S, Brage S, Wareham NJ, McDermott CJ, *et al.* Long-term physical activity: An exogenous risk factor for sporadic amyotrophic lateral sclerosis? *Amyotroph Lateral Scler Front Degener* 2016;17:377-84.
40. Beghi E, Logroscino G, Chiò A, Hardiman O, Millul A, Mitchell D, *et al.* Amyotrophic lateral sclerosis, physical exercise, trauma and sports: Results of a population-based pilot case-control study. *Amyotroph Lateral Scler* 2010;11:289-92.
41. Patel BP, Hamadeh MJ. Nutritional and exercise-based interventions in the treatment of amyotrophic lateral sclerosis. *Clin Nutr* 2009;28:604-17.
42. Carreras I, Yuruker S, Aytan N, Hossain L, Choi JK, Jenkins BG,

- et al.* Moderate exercise delays the motor performance decline in a transgenic model of ALS. *Brain Res* 2010;1313:192201.
43. Chio A, Moglia C, Canosa A, Manera U, D'Ovidio F, Vasta R, *et al.* ALS phenotype is influenced by age, sex, and genetics: A populationbased study. *Neurology* 2020;94:8.
44. AlChalabi A, Hardiman O, Kiernan MC, Chiò A, RixBrooks B, van den Berg LH. Amyotrophic lateral sclerosis: Moving towards a new classification system. *Lancet Neurol* 2016;15:118294.