UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Risk factors for Amyotrophic Lateral Sclerosis (ALS): the effect of high-intensity sport

Ana Rita Oliveira Henriques

Mestrado em Bioestatística

Trabalho de projeto orientado por:
Professor Doutor Ruy Ribeiro
Professora Doutora Marília Antunes

2020

Agradecimentos

À Professora Doutora Marília Antunes e ao Professor Doutor Ruy Ribeiro, pela disponibilidade, apoio e orientação fundamental para a realização deste trabalho, um muito obrigada.

Ao Professor Doutor Mamede de Carvalho e à Professora Doutora Marta Gormicho pela disponibilidade e gentileza da partilha dos dados trabalhados nesta tese.

A todos aqueles que de algum modo contribuíram para a realização deste trabalho expresso aqui o meu reconhecimento e gratidão.

Por fim, um especial agradecimento à minha família e amigos mais próximos pelo amor e apoio incondicional.

Resumo Alargado

A Esclerose Lateral Amiotrófica (ELA) é uma doença neurodegenerativa rara que envolve principalmente os neurónios motores (diretamente associados a movimentos voluntários). Atualmente, não existe cura para a ELA, aproximadamente 50% dos pacientes com ELA morrem dentro de 30 meses após o início dos sintomas, geralmente devido a insuficiência respiratória e suas complicações, enquanto cerca de 10% dos pacientes podem sobreviver por mais de uma década (National Institute of Neurological Disorders and Stroke NIH, n.d.).

A maioria dos casos de ELA é esporádica, sem histórico familiar da doença. No entanto, 5 a 10% dos casos têm uma componente genética associada - uma mutação do gene C9ORF72, responsável pela codificação da proteína C9ORF72, mais comumente associada à ELA. Outra mutação sobre a qual há referências de ligação com a ELA ocorre no gene SOD1, que em conjunto com a mutação do gene C9ORF72, explica os 87% da ELA familiar. (Renton et al., 2011).

A idade e o género, são fatores de risco já bem estabelecidos. Para além destes, outros fatores como o tabagismo ou a prática de atividade física intensa, são apontados como potencias fatores de risco para a doença, embora os resultados de trabalhos anteriores sejam contraditórios (Ingre, Ross, Phiel, Kamel, & F, 2015). Ainda assim, estudos recentes indicam que as modalidades de contacto nos níveis mais altos de intensidade, que combinam atividade física intensa e aumento do risco de trauma repetitivo na cabeça e na coluna cervical (Blecher et al., 2019), são um fator de risco para o aparecimento de doenças neurodegenerativas.

Esta tese, baseou-se num estudo caso-controlo da ELA, e teve como principais objetivos avaliar a relação entre a prática das modalidades de contacto - boxe, rugby, hóquei e futebol e a ocorrência de ELA, bem como a existência de potenciais fatores que interfiram nessa relação, como sexo, idade e hábitos tabágicos. Optou-se por selecionar apenas as modalidades que envolviam trauma repetitivo ao nível da cabeça e coluna cervical, partindo do pressuposto, apresentado por (Blecher et al., 2019). A análise estratificada em estudos de caso-controlo e modelos de regressão logística, foram as metodologias escolhidas para atingir este objetivo. Os resultados obtidos pelas duas metodologias foram comparados.

O segundo objetivo, foi avaliar a relação entre a idade de diagnóstico precoce e a prática de futebol de alta intensidade. Para testar a existência de diferenças na idade do diagnóstico, considerando os diferentes níveis de intensidade com que o futebol era praticado, foi aplicada a análise de variância a um fator.

Foram também definidos objetivos secundários: i) avaliar a relação entre a idade do diagnóstico e a idade dos primeiros sintomas e o tipo de início da doença, através do modelo linear clássico. ii) avaliar a relação entre os diferentes valores de *ALS Fuctional Rate of Decay* (ALSFSR.R) e um conjunto de variáveis como sexo, idade dos primeiros sintomas e outras através da mesma metodologia. iii) avaliar a existência de diferenças entre os diferentes níveis de progressão por meio de regressão logística multinominal. Quando uma ou mais suposições não foram atendidas, recorreu-se à aplicação da transformação logarítmica.

Para todas as análises estatísticas realizadas, com exceção da análise estratificada, foi utilizado o software R com a interface RStudio Versão 1.1.463 -c © 2009-2018 RStudio, Inc. A análise estratificada foi realizada com o software baseado na Web Open-Epi. Foi considerado o nível de significância $\alpha = 5\%$.

O conjunto de dados utilizado compreende 2247 participantes, 1326 pacientes e 921 controlos de várias nacionalidades Europeias que responderam a um questionário padronizado, criado em 2015 como parte do projeto OnWebDuals. O projeto OnWebDuals pretende recolher dados de pacientes com ELA de diferentes locais europeus, para construir uma ontologia de domínio da ELA e implementá-la num grande banco de dados web-europeu. Fazem parte desta base de dados variáveis sociodemográficas como por exemplo género e idade, variáveis clínicas como tipo de início da doença, a idade dos primeiros sintomas e ainda uma classe de variáveis exclusivamente relacionadas com o desporto, tendo em conta um dos objetivos principais desta tese.

Encontrou-se evidência para o nível de significância $\alpha = 5\%$ para que género e idade (categorizada) atuem como modificadores de efeito na relação entre a prática de modalidades de contacto e a ELA. Contudo, após maior reflexão, é levantada a possibilidade destes resultados estarem relacionados com o número reduzido de mulheres a praticar modalidades de contacto, o que é natural, considerando que nesta base de dados a idade média das mulheres é de 63.5 anos. Nesta geração, o número de mulheres que praticavam atividade física, era muito reduzido. Desta forma, foi construido um modelo onde se considerou exclusivamente a população masculina.

Quando considerado apenas o efeito da prática de modalidades de contacto, esta surge como fator de risco para a doença, a chance de vir a desenvolver a doença é 1.765 vezes superior nos praticantes, relativamente aos não praticantes. Houve também interesse em avaliar o efeito de outras variávéis consideradas fatores de risco, ou potenciais fatores de risco, para a doença - idade e hábitos de fumo representados pela carga tabágica (TE). Foram testados diferentes modelos, chegando ao modelo final.

O modelo final, constituído exclusivamente pela população masculina tem como variáveis - prática de modalidades de contacto (Contact), idade, carga tabágica (TE) e interação entre a prática de modalidades de contacto e a carga tabágica (Contact $\times$ TE). Considerando o objetivo principal - avaliar a relação entre a prática de modalidades de contacto e a ELA, pode dizer-se que a chance de vir a desenvolver a doença é 1.3 (IC a 95%:0.867;2.124) vezes maior nos praticantes relativamente aos não praticantes nos participantes com carga tabágica nula (TE=0). Quanto à interação entre a prática de modalidades de contacto e a carga tabágica, os resultados mostraram-se estatisticamente significativos, verifica-se um aumento de aproximadamente 3% na chance de vir a desenvolver a doença nos praticantes de modalidade de contacto, quando comparados com os não praticantes para duas cargas homogéneas que diferem em 1 maço/ano.

Relativamente ao segundo objetivo, para além de se avaliar a relação entre a prática de futebol de alta intensidade e idade de diagnóstico precoce, considerando todos os participantes dos 4 países Europeus (Alemanha, Polónia, Portugal e Turquia), procedeu-se à mesma avaliação, mas aplicada exclusivamente aos participantes portugueses, que nos interessava especialmente. Quando considerada apenas a população portuguesa pode dizer-se que a idade de diagnóstico difere significativamente entre praticantes de futebol de alta intensidade e não praticantes. Contudo, quando considerada toda a população esta diferença não se verifica.

Relativamente aos objetivos secundários, por cada ano mais na idade de ocorrência dos primeiros sintomas, verificou-se uma diminuição de aproximadamente 0.7% no atraso do diagnóstico. Comparando pacientes com início bulbar relativamente a pacientes com início num dos membros, pode dizer-se que o atraso no diagnóstico é aproximadamente 30% menor para os pacientes com início bulbar relativamente aos com início nos membros. Considerando os diferentes grupos de progressão, verificou-se que grupos

de progressão mais lentos correspondem a um aumento no atraso de diagnóstico.

Apesar de ainda não existirem muitos estudos nesta área, este é um dos maiores considerando o tamanho da amostra. Os resultados obtidos e agora apresentados, estão de acordo com alguns trabalhos publicados anteriormente. Dado o tamanho da nossa amostra, estes resultados, são atuais e importantes para validar outros trabalhos semelhantes que têm vindo a ser desenvolvidos.

No futuro, será importante, a realização de um estudo semelhante, para confirmar o efeito de modalidades de alta intensidade associadas a um trauma repetitivo a nível da cabeça e coluna cervical como um fator de risco para a ELA. A sua aplicação a uma nova geração de homens e mulheres, em que as diferenças relativas à prática de atividade física não sejam tão acentuadas, permitirá obter resultados mais ajustados à realidade atual.

Citando a emblemática afirmação de (Box, 1979), ”todos os modelos estão errados, mas alguns são úteis”. Os resultados agora apresentados, não podem ser assumidos como lei universal mas reforçam a necessidade de mais investigação nesta área, contribuindo para um melhor conhecimento da doença.

A confirmação destes resultados deverá proporcionar uma maior consciencialização das federações desportivas e dos atletas para este problema e promover a discussão de estratégias que de alguma forma tentem minimizar os efeitos da prática destas atividades na saúde dos atletas.

Palavras-chave: Esclorose Lateral Amiotrófica, Modalidades de Contacto e Modelo Linear Generalizado.

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a rare neurodegenerative disease that primarily involves motor neurons (directly associated with voluntary movements). Currently there is no cure for ALS, approximately 50% of ALS patients die within 30 months of symptom onset, usually due to respiratory failure and its complications, while about 10% of patients may survive more than a decade (National Institute of Neurological Disorders and Stroke NIH, n.d.).

Vigorous, high-intensity physical activity has been identified as a risk factor for ALS; the practice of modalities prone to repetitive injuries of the cervical spine and head are pointed as possible risk factors for ALS when associated with vigorous practice (Blecher et al., 2019).

This report has two main objectives: i) evaluate the relationship between the practice of contact modalities - boxing, hockey, football and rugby and ALS and the existence of variables that can interfere in this relation: age, gender and smoking habits. ii) evaluate the relationship between early diagnosis and football practice with different levels of intensity. In complement, variables associated with late diagnosis and different levels of disease progression were also evaluated.

The study included 2247 individuals - 1326 patients and 921 controls. To answer the first objective, binary logistic regression models were used, the response variable was the presence or absence of ALS. The explanatory variables considered were the practice of contact modalities, the study main variable and age, gender and smoking (as packs/year). Stratified analysis in case control studies was also applied, with the same objective. The results obtained by both methodologies were compared. For other purposes, classical linear model and its extensions were used. The significance level $\alpha = 5\%$ was considered.

The results suggested the existence of evidence indicating gender and age as effect modifiers in the relationship between ALS and contact sports. The small number of female practicing contact modalities and their high average age, may have conditioned the results, so special attention was paid to it and only the male population was considered. A relationship between the practice of contact modality and ALS, was found. For the second objective, considering only the Portuguese population, statistically significant differences were found in the age of diagnosis between high intensity practitioners and practitioners with low intensity level or non-practitioners. However, this difference was not verified, when considering the whole population. Limbs onset and age of early diagnosis are associated with an increase in diagnostic delay. Considering the different progression groups, the slower progression correspond to an increase in the delay of the diagnosis.

Although this is still a poorly studied area, as there is a growing concern around the effects of physical activity on this disease, the results presented now, are important and in accordance with similar previous works, and supports and validate them. Contact × TE interaction, was statistically significant. As far as known, this interaction was not yet identified and there are no previous published results showing it. Future similar studies are needed to include a new generation of men and women where differences regarding the practice of physical activity are not so pronounced.

Keywords: Amyotrophic Lateral Sclerosis, Contact Sports, Generalized Linear Models.

Index

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
ALS	Amyotrophic Lateral Sclerosis
ALSFRS	ALS Functional Rating Scale
ALSFRS-R	ALS Functional Rating Scale Revised
ANOVA	Analysis of Variance
AUC	Area Under the Curve
CI	Confidence interval
GLM	Generalized Linear Models
MLE	Maximum Likelihood Estimators
MSE	Mean Squared Error
OR	Odds-Ratio
ONWebDUALS	OnWebDuals-Web-based ontology database
RCS	Restricted Cubic Splines
ROC	Receiver Operating Characteristic
SSE	Sum of Squares of the Residuals
SSF	Sum of Squares associated to the Factor

Chapter 1

Introduction

This report was based on a case-control study for Amyotrophic Lateral Sclerosis (ALS). In this chapter a brief presentation of the disease and its main risk factors is made. Objectives and a summary description of the following chapters are also presented bellow.

ALS is a rare neurodegenerative disease, involving mainly motor neurons (directly associated with voluntary movements). It is a progressive disease, meaning that functional impairment worsens over time. Currently, there is no cure for ALS. The most frequent symptoms are muscle fasciculations and limb weakness, spreading to other regions of the body. Approximately 50% of ALS patients die within 30 months of onset of symptoms, often due to respiratory failure and its complications, while about 10% of patients can survive for more than a decade (National Institute of Neurological Disorders and Stroke NIH, n.d.)

In Europe the disease has a more common incidence in men than in women, usually peaking in ages between 50 and 75 years (Logroscino et al., 2010) and generally affects 2-3 people in 100,000 with an overall life-time risk of developing the disease of 1:400 (Hardiman, van den Berg, & Kiernan, 2011). The first sign of ALS may appear in one of the lower or upper limbs; when symptoms begin in the arms or legs, it is referred to as "limb onset ALS", in other cases, the first symptoms are characterized by speech difficulties or swallowing problems, termed "bulbar onset ALS" (National Institute of Neurological Disorders and Stroke NIH, n.d.), this two signs are the most common ones. In addition to those, patients might have other forms of onset, including mixed onset (spinal and bulbar), thoracic onset, thoracic onset or respiratory symptoms (Longinetti & Fang, 2019). Respiratory onset and suffer from breathing difficulties even when standing are generally visible symptoms in more advanced ALS patients (Gautier et al., 2010).

Most cases of ALS are sporadic, with no family history of the disease. However, 5 to 10% of the cases have an associated genetic component, in which the individual inherits the disease from the parents.

The C9ORF7 gene, which encodes the C9ORF72 protein is the most commonly mutated gene in ALS. This mutation causes a noncoding stretch of 30 nucleotides to be expanded hundreds, even thousands, of times and is a major cause of familial ALS (Renton et al., 2011). Along with the SOD1 mutation, they explain 87 % of familial ALS, being pointed as the most common cause known for this disease (Renton et al., 2011).

In addition to aging and gender, well-established risk factors, other proposed factors include smoking and heavy physical activity. Despite the existence of some studies on the relationship between the disease and these factors, the results are still inconsistent so it is important to have more in-depth investigation on this field.

Cigarette smoking might contribute to the risk of ALS either by a direct neurotoxic effect on motor neurons or by increasing oxidative stress. Although there are recent studies that point to smoking as the major environmental factor related to ALS, there have been relatively few studies and results have been conflicting (Wang, Reilly, Weisskopf, Kolonel, & Ascherio, 2012).

Vigorous physical activity has been pointed out as a risk factor. This association is biologically plausible, because vigorous exercise may induce oxidative stress and glutamate excitotoxicity (Pupillo et al., 2014).

The practice of sports prone to repetitive head and cervical spine injuries, such as football, is pointed as potential risk factors for the onset of ALS, when associated with a vigorous practice. Yet analysis of how the level of competitiveness (professional *vs* nonprofessional) or the type of sport (contact *vs* non-contact) affect this risk remains unanswered (Blecher et al., 2019) .

Another factor that has been cited as a cause for ALS is cyanobacteria toxin $\beta - N - methylamino - l - alanine(BMAA)$(Cox, Kostrzewa, & Guillemin, 2018). Combined with multiple mechanisms of BMAA neurotoxicity, particularly to vulnerable subpopulations of motor neurons, its production by cyanobacteria has significantly increased interest in investigating exposure to this non-protein amino acid as a possible risk factor for other forms of neurodegenerative illness all around the world (Cox et al., 2018). As football is practiced on lawns, where cyanobacteria are present, this factor usually appears associated with the practice of this activity.

It is still difficult to understand disease progression, and what makes a patient progress faster or slower.

The ALS Functional Rating Scale (ALSFRS) is a standard test used by physicians to estimate the outcome of a treatment or the progression of the disease. Although very popular, this scale has only a small respiratory component. Given that respiratory failure is the most common cause of death in ALS patients, the ALS functional rating scale revised (ALSFRS-R) was proposed (Cedarbaum et al., 1999).

This monitoring is usually performed using a standard questionnaire consisting of 12 expert questions to assess motor function, bulbar symptoms, and respiratory capacity in ALS patients. Each question is rated on a scale of (0 to 4), where 0 represents total performance loss and 4 represents normal performance. The sum of all questions generates a score of up to 48, usually called the ALSFRS-R score, which represents the patient's current state (Cedarbaum et al., 1999). By measuring the change in ALSFRS-R over time, we can obtain an estimation of how the disease is progressing and infer about the survival of the patient (Pires, Gromicho, Pinto, Carvalho, & Madeira, 2019).

According to (Pires et al., 2019), patients can be stratified into separate groups by computing their progression rates. The progression rate (*ProgressionRate*) is an attribute measuring how fast ALS is progressing in a patient, and is based on the recorded values in the patient's ALSFRS-R test results, trough the equation:

$$ProgressionRate = \frac{48 - ALSFRSR_{1^{st}Visit}}{\triangle t_{1^{st}Symptoms;1^{st}Visit}} \tag{1.1}$$

where 48 is the maximum score for the ALSFRS-R scale, $ALSFRSR_{1^{st}Visit}$ is the ALSFRS-R score of a given patient in the first appointment (diagnosis) and $\triangle t_{1^{st}Symptoms;1^{st}Visit}$ is the time in months between the first symptoms and the first appointment.

Considering a patient who gets the maximum score on the questionnaire, *ProgressionRate* will be null. By taking into account *ProgressionRate* values, the disease progression levels can be divided in 3 groups: slow values of *ProgressionRate* between 0 and 0.5, neutral for values between 0.5 and 1.5 and finally fast for values greater than 1.5.

Imagine the following example, where a patient who obtained 44 points in the questionnaire

ALSFRS-R score, which shows signs of disease at 25 months, based on the expression (1.1) will obtain an *ProgressionRate* of 0.16.

$$ProgressionRate = \frac{48-44}{25} = 0.16$$

The ONWebDUALS project

The participants analysed in this thesis were selected from the OnWebDuals-Web-based ontology database project to understand ALS. The ONWebDUALS is a European project, coordinated by the ALS center in Lisbon, with support of the Joint EU Program for Neurodegenerative Diseases (JPND). ONWebDUALS aims to combine standardized phenotypes to identify risks and relevant prognostic factors in ALS. The main expected result of the ONWebDUALS project is to lessen the social impact of ALS in Europe.

This is a case-control study, to retrospectively evaluate the effect of physical activity and multiple other factors that may influence the occurrence of ALS. Two groups were selected: one of patients with ALS and another of individuals without the disease (control group). Participants were asked to answer a questionnaire about their lifestyle, sports habits, smoking habits and family history of illness. For patients, the database used for the current study, also contains information about the disease degree and onset type. The OnWebDUALS project was approved by the relevant Ethical Committees. All data analyzed in the current study is completely anonymous without possibility of subject identification.

Objectives

Two main objectives were defined. The first one was to evaluate the relationship between the practice of contact modalities-boxing, rugby, hockey and football and the occurrence of ALS, as well as the existence of potential interfering factors in this relationship, such as gender, age and smoking habits. Stratified analysis in case-control studies and logistic regression models were used. The results obtained by the two methodologies were compared.

The second objective was to evaluate the relationship between the early age of diagnosis and the practice of high intensity football. To test the existence of differences in the age of diagnosis considering the different levels of intensity with which football was practiced, analysis of variance was applied.

Secondary objectives were also defined. The evaluation of the relationship between the age of diagnosis and the age of the first symptoms and onset type through the classic linear model was considered relevant. The same methodology was used to evaluate another similar secondary objective, the relation between different values of *ProgressionRate* and a set of variables such as gender, age of first symptoms and others.

As is well known, the classical linear model is very strict in relation to its assumptions, therefore, if one or more of the assumptions are not met, the logarithmic transformation was applied. If the logarithmic transformation is not effective, an alternative to apply is the e gamma distribution and the log link function.

The existence of differences between the different levels of progression was also assessed through multinomial logistic regression.

Thesis structure

Chapter 2 presents, describes and introduces the basic theoretical ideas and the procedures of the main statistical analysis methods used - multiple linear regression analysis, logistic regression analysis and stratified analysis in case-control studies. Chapter 3 presents the procedures related and applied to the database, providing a description of the used variables and the new variables created from them. After the application of all the methodologies previously mentioned, the results obtained are presented in Chapter 4. In Chapter 5, discussion and conclusion are presented. For the discussion, a comparison with the results obtained in previous studies, was included. Limitations of the used methodology are also discussed. Finally, a summary with the main conclusions of the developed work is presented.

Chapter 2

Methods

In this chapter a general presentation of the Generalized Linear Models (GLM) is made, specifying the models used throughout the thesis. Reference to stratified analysis is also made.

For all statistical analyses performed, with the exception of stratified analysis, the software R with the interface RStudio Version 1.1.463 - © 2009-2018 RStudio, Inc, was used. The stratified analysis was performed using the web-based software Open-Epi (www.OpenEpi.com) a free, Web-based, open-source, operating system-independent series of programs designed for use in public health and medicine - for training or practice - that provide a number of epidemiologic and statistical tools (Sullivan, n.d.).

2.1 Generalized Linear Models

GLM were the basis of the whole methodology of this report, these are a generalization of the classical linear model. This generalization is mainly related to the fact that the distribution associated with the response variable is not restricted to Normal and can be any distribution belonging to the exponential family of distributions.

Examples of these models are: analysis of variance and covariance analysis, logistic regression, poisson regression, log-linear models, probit regression and others.

The distribution function associated with the response variable in a GLM is part of the exponential family, defined as:

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - \mathrm{b}(\theta)}{\mathrm{a}\phi} + \mathrm{c}(y,\phi)\right\} \tag{2.1}$$

where θ is a canonic form for the location parameter, ϕ is the dispersion parameter and a(.), b(.) and c(.) are known real functions.

Several frequently used distributions, such as the Normal, Poisson and Binomial distributions, belong to the exponential family of distributions.

Here, special attention was given to the binomial family, the logistic regression was choosen to meet the main objectives. Later, this models will be presented with more detail, but now, the focus goes to the classic linear model, which has also played an important role for the development of this work.

2. METHODS

2.1.1 Classical Linear Model

In a linear model, the dependent variable Y_i $(i = 1,2,..,n)$ is modeled as a linear fuction of (p) independent variables $x_1, x_2, ..., x_p$ as

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \tag{2.2}$$

The classical linear regression model assumptions are: I) Independence: the values of Y_i are statistically independent of each other; II) Linearity: the expected value of Y_i, is a linear function of $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ III) Normality: the error terms follow a Normal distribution $\varepsilon_i \frown N(0, \sigma^2)$ and IV) Homoscedasticity: the errors have the same variance σ^2, independent of the values of X.

Different diagnostic tools have been developed to check these assumptions. To evaluate the assumption of linearity (II), equality of variance (homoscedasticity) (IV) and error normality (III) in linear regression, the choice was to plot the residuals. For the first two assumptions residuals vs fitted plot were used. For the third assumption QQ plot of residuals was used. The assumption of independence is verified by the collection protocol.

If one or more of the linear regression assumptions is violated, the results of the analysis may be incorrect or misleading. Therefore, it is necessary to look for alternative models as there are a set of transformations when one or more assumptions are not verified.

Log transformations are one of the most commonly used transformations. In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed, more specifically, if a variable Y follows a log-normal distribution, then we have that $\ln(Y)$ follows a normal distribution with a mean= μ and a variance= σ^2 (Cornell Statistical Consulting Unit, 2012). In terms of the original variable Y, the most important properties of the log-normal distribution are: the expected value of Y is $E(y) = \exp(\mu + \frac{1}{2}\sigma^2)$; median or geometric mean of variable $Y = \exp(\mu)$ and variance of variable y is $(\exp(\sigma^2) - 1)) \times \exp(2\mu + \sigma^2)$.

Considering these properties:

$\ln[E(Y)] \neq E[\ln(Y)] \Leftrightarrow E(Y) \neq \exp(E[\ln(Y)])$

$\ln[Median(Y)] = Median[\ln(Y)]$

It is easily concluded that, $E(Y) \neq \exp(\beta^x)$, exponentiation cannot be used to interpret the model coefficients. However, $Median(Y) = \exp(x\beta)$.

Being the median a statistical measure of location, and the logarithmic function, a growing injective function, the median can be used to interpret the model. Models for a log transformed outcome yield inferences to the arithmetic mean in log scale, which is equivalent to the geometric mean in original scale.

Considering the simple linear model there are three possible combinations of transformations involving logarithms: **1)** the linear-log model i.e., the independent variable is log transformed, **2)** the log-linear model i.e., the dependent variable is log transformed and **3)** the log-log model, both dependent and independent variables are log transformed .

Considering the three hypotheses, a linear model can be expressed with the following equations:

1- Linear-log model

$$Y_i = \beta_0 + \beta_1 \ln(x_{i1}) + ... + \beta_p x_{ip} + \varepsilon_i \tag{2.3}$$

2- Log-linear model

$$\ln(Y_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \tag{2.4}$$

3- Log-log model

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(x_{i1}) + ... + \beta_p x_{ip} + \varepsilon_i \tag{2.5}$$

Below the interpretation in a simple linear regression setting when the dependent variable or both variables are log transformed, will be explored.

Consider an example that studies the relationship between height and weight. People's weights tend to have a higher variance for taller people, so it is quite reasonable to take log of weight when you are fitting a linear regression model. This example was adapted from Cornell Statistical Consulting Unit (2012).

Suppose the dependent variable is log-transformed, and the regression is estimated as follows:

$$E[\ln(Y)] = \beta_0 + \beta_1 Age$$

$$Median(y) = \exp(\beta_0) \times \exp(\beta_1 Age)$$

$$\frac{Median(y) \mid Age = x+1}{Median(y) \mid Age = x} = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1(x))} = \exp(\beta_1) \tag{2.6}$$

If:

$$\ln(\widehat{ALSFRS.R}) = 2.14 + 0.005 Age \tag{2.7}$$

The estimated coefficient of the Age variable is β_1 =0.005, so that means that an increase of one-unit in the Age would result in (exp(0.005)) an increase of approximately 0.5% in the ALSFRS.R.

If both the dependent variable and independent variable are log-transformed, the fitted regression is:

$$E[\ln(Y)] = \beta_0 + \beta_1(\ln(Age))$$

$$Median(y) = \exp(\beta_0) \times \exp(\beta_1(\ln(Age)))$$

These cases have to be thought mathematically $\ln(x) + \ln(y) = \ln(x \times y)$, therefore the multiplying factor can be considered

$$\frac{Median(y) \mid Age = x \times 1.01}{Median(y) \mid Age = x \times 1} = \frac{\exp(\beta_0) \times (1.01x)^{\beta_1}}{\exp(\beta_0) \times (x)^{\beta_1}} = (1.01)^{\beta_1} \tag{2.8}$$

Suppose:

$$\ln(\widehat{ALSFRS.R}) = 1.69 + 0.2\ln(Age)$$

Here $\beta_1 = 0.2$. That means that a one percent change in Age is associated with an increase of approximately 0.2% $(1.01^{0.2})$ in the ALSFRS.R.

2. METHODS

One-way ANOVA

To evaluate the relationship between a quantitative response variable and a factor, the one-way ANOVA is generally used.

The one-way ANOVA model is written in the following manner:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{2.9}$$

where μ represent the general level of the response variable in the population; α_i represent the effect of level i or the effect of treatment i and ε_{ij} is an error variable that follows the usual linear-model assumptions—that is, the ε_{ij} are independent and normally distributed with mean zero and equal variance.

By analyzing the general equation (2.9) it is easy to identify the presence of a set of k equations, serving the observations of each level of the factor:

- for the n_1 observations made at the level $i = 1$,

$$Y_{1j} = \mu + \alpha_1 + \varepsilon_{1j},$$

- for the n_2 observations made at the level $i = 2$,

$$Y_{2j} = \mu + \alpha_2 + \varepsilon_{2j},$$

- ...

- for the n_k observations made at the level $i = k$,

$$Y_{kj} = \mu + \alpha_k + \varepsilon_{kj}.$$

This set of k equations can be written as a single equation, which is the equation of a linear model:

$$Y_{ij} = \mu + \alpha_1 I_{1ij} + \alpha_2 I_{2ij} + ... + \alpha_K I_{kij} + \varepsilon_{ij},$$

where $I_{mij} = 1$ if the observation belongs to level $i = m$, and $I_{mij} = 0$ otherwise. The variables I_{mij} are indicator variables of each factor level.

The model has an excess of parameters. Just note that there is a dummy variable for each of the k categories of the treatment.

In order to reduce the parameters, there are some possible solutions, such as take α_1 =0.

So, assuming that $\alpha_1 = 0$, then $\mu = \mu_1$:

$$\mu_1 = \mu, \forall_j = 1, ..., n_1$$

$$\mu_2 = \mu_1 + \alpha_2, \forall_j = 1, ..., n_2$$

$$\mu_3 = \mu_1 + \alpha_3, \forall_j = 1, ..., n_3$$

$$...$$

$$\mu_k = \mu_1 + \alpha_k, \forall_j = 1, ..., n_k$$

In the model for a one-way ANOVA, each α_i, ($i > 1$), represents the increment that turns the average of the first level into the average of the level i:

$$\begin{array}{rcl} \alpha_1 & = & 0 \\ \alpha_2 & = & \mu 2 - \mu_1 \\ \alpha_3 & = & \mu 3 - \mu_1 \\ \vdots & \vdots & \vdots \\ \alpha_k & = & \mu_k - \mu_1 \end{array}$$

The equality of all population means,$\mu_1 = ... = \mu_k$, is equivalent to state that all level effects are null: $\alpha_i = 0, \forall_i$.

From data presented above,the natural estimators are:

$$\hat{\mu}_1 = \overline{Y}_1$$

and

$$\hat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{1.},$$

where

$$\overline{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

So, it follows that any observation has a fitted value:

$$\hat{Y}_{ij} = \hat{\mu}_i = \hat{\mu}_1 + \hat{\alpha}_i = \overline{Y}_i$$

To test factor effects, the hypothesis that none of the factor levels affect the mean of the variable or, equivalently, that all treatments produce the same effect, corresponds to the hypothesis

$$H_0 : \alpha_2 = \alpha_3 = ... = \alpha_k = 0,$$

equivalent to

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k = (\mu).$$

The null hypothesis of no differences among population group means is tested by the F-test.

Under the one-way ANOVA model,to test

$$H_0 : \alpha_i = 0 \quad \forall_{i=2,...,k} \quad vs \quad H_1 : \exists_{i=2,...,k} : \alpha_i \neq 0,$$

the test statistic is

$$F = \frac{MSF}{MSE} \frown F(k-1, n-k), \quad \text{under} \quad H_0. \tag{2.10}$$

At a significance level α,

$$\text{reject } H_0 \text{ if } F_{obs} > F_{1-\alpha};(k-1, n-k).$$

In this context, the Sums of Squares and Mean Squares also have specific formulas in one-way ANOVA:

The sum of squares of the residuals (**SSE**) is given by:

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2, \tag{2.11}$$

where $\mathrm{S}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y}_{i.})^2$is the sample variance of the $n - i$ observation of Y at the level i of the factor.

SSE measures the variability in the "inside"of the k levels.

The sum of squares associated to the factor (**SSF**) is given by:

$$SSF = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\widehat{Y}_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2 \tag{2.12}$$

It can also be written

$$SSF = \sum_{i=1}^{k} n_i(\widehat{Y}_{ij} - \overline{Y}_{..})^2,$$

where $\bar{Y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij})$ is the average of all the observations (n).

SSF it is a measure of between-groups variability.

This information can be presented in an ANOVA table.

Table 2.1: General One-Way Analysis-of-Variance Table

Source	d.f	SS	MS	F$_{obs}$
Factor	$k-1$	SSF=$\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	MSF= $\frac{SSF}{k-1}$	$\frac{MSF}{MSE}$
Residuals	$n-k$	SSE=$\sum_{i=1}^{k}(n_i - 1)S_i^2$	MSE= $\frac{SSE}{n-k}$	
Total	$n-1$	SST= $(n-1)s_y^2$	-	-

Multiple Comparasion

By rejecting H_0 (in favor of H_1), the question of what levels of the factor whose means differ from one another remains open (except when $k = 2$), the number of comparisons to be made is considerable, since all possibilities have to be taken into account.

Suppose that $k = 3$. Rejection of H_0 may be due to:

- $\mu_1 = \mu_2 \neq \mu_3$, this is, $\alpha_2 = 0$; $\alpha_3 \neq 0$
- $\mu_1 = \mu_3 \neq \mu_2$, this is, $\alpha_3 = 0$; $\alpha_2 \neq 0$
- $\mu_1 \neq \mu_2 = \mu_3$, this is, $\alpha_2 = \alpha_3 \neq 0$
- $\mu_1 \neq \mu_2 \neq \mu_3$, this is, $\alpha_2 \neq \alpha_3, \alpha_2, \alpha_3 \neq 0$

The question that arises is how to decide between these different alternatives, this is where the multiple comparison tests will come in.

The most commonly used test for multiple comparisons for balanced groups is the Tukey test, or the Tukey-Kramer test, as an alternative for unbalanced groups. Tukey test is based on the following result (Tukey Distribution):

Let $\{W_i\}_{i=1}^{k}$ be independent random variables with normal distribution with the same parameters:

$$\mathrm{W}_i \frown N(\mu_w, \sigma_w^2)\ \forall i = 1..., k.$$

-Let $R_W = \max_i W_i - \min_i W_i$ be the range of the total sample.

-Let S_w^2 be an estimator of the common variance, σ_w^2, such that $\frac{vS_w^2}{\sigma_w^2} \frown \chi_v^2$.

- Let S_w and R_w be independent r.v. .

Then, the Studentized range, $\frac{R_w}{S_w}$, follows a Tukey distribution dependent on two parameters: K and v.

In a one-way ANOVA, we have

$$\overline{Y}_{i.} \frown N\left(\mu_i, \frac{\sigma^2}{n_i}\right) \quad \Leftrightarrow \quad \overline{Y}_{i.} - \mu_i \frown N\left(0, \frac{\sigma^2}{n_i}\right)$$

If the design is balanced, that is, $n_1 = n_2 = n_k (= n_c)$, the k differences $\overline{Y}_{i.} - \mu_i$ will follow the same distribution, $N(0, \frac{\sigma^2}{n_c})$, and the variables will be W_i r.v

An estimator of the common variance, $\frac{\sigma^2}{n_c}$ is given by $S_w^2 = \frac{MSE}{n_c}$ and given that the remaining conditions are met,

$$\frac{R_w}{S_w} = \frac{\max_i(\overline{Y}_{i.} - \mu_i) - \min_j(\overline{Y}_{j.} - \mu_j)}{\sqrt{\frac{MSE}{n_C}}} \tag{2.13}$$

follows a Tukey distribution with parameters k and $n-k$.

Thus, the hypothesis under test for balanced designs is,

$$H_0 : \mu_i = \mu_j, \forall i, j \quad vs \quad H_1 : \exists i, j : \mu_i \neq \mu_j$$

the test statistic is

$$\frac{R_w}{S_w} \frown Tukey_{(k,n-k)} \, \text{under} H_0.$$

At a significance level α, reject $H_0 : \mu_i = \mu_j$ if, for any pair (i, j)

$$\mid \overline{Y}_{i.} - \overline{Y}_{j.} \mid > q_{1-\alpha}(K, n-k)\sqrt{\frac{MSE}{n_c}}$$

A confidence interval for $\mu_i - \mu_j$ with a confidence level of $100(1-\alpha)\%$ is

$$\left(\overline{y}_{i.} - \overline{y}_{j.} - q_\alpha(k, n-k)\sqrt{\frac{MSE}{n_c}}, \overline{y}_{i.} - \overline{y}_{j.} + q_\alpha(k, n-k)\sqrt{\frac{MSE}{n_c}}\right) \tag{2.14}$$

When the one-way ANOVA sample design is not balanced (that is, when the n_i are not all equal), the Tukey tests / CIs now stated are not, strictly speaking, valid (Ito, 1980).

As mentioned above, the Tukey-Kramer method is an alternative to the Tukey test. Unlike the Tukey test, Tukey-Kramer test allows unequal sample sizes between treatments and is therefore more frequent in unbalanced designs.

The Tukey-Kramer test has the same goal as the Turkey test so the hypotheses under test are the same.

Tukey-Kramer test declares two significantly different means if the absolute value of their sample differences exceeds

$$q_{1-\alpha}(k,n-k)\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{2.15}$$

and the confidence interval for $i \neq j$ is

$$\left(\bar{y}_{i.}-\bar{y}_{j.}-q_{1-\alpha}(k,n-k)\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},\bar{y}_{i.}-\bar{y}_{j.}+q_{1-\alpha}(k,n-k)\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\right) \tag{2.16}$$

Robustness of ANOVA

ANOVA is a bit more flexible regarding non-compilance with assumptions when compared to linear regression. There have been some studies of the robustnes of the effects of nonnormality and/ or heterocedasticity on the ANOVA F-test (Ito, 1980).

The F-test is found to be remarkably insensitive to general nonnormality. If the group sample sizes are equal, the F-test is not very sensitive to heterogeneity of variance from group to group. However with unequal groups much greater effects can occur on the F-test when variances are different from group to group (Ito, 1980).

The residuals plot was used to verify the assumptions, residuals vs fitted plot was used for test homoscedasticity and QQplot for normality.

2.1.2 Stratified Analysis in case-control studies and Logistic Regression

A description of the methodology used to meet the main objective - to assess the relationship between contact modalities practice and ALS, is presented here.

Multivariable models are widely used in epidemiology to describe the relationship between a result (dependent variable or response) and a simultaneous set of explanatory variables (predictor or independent). This allows a confounding adjustment and evaluation of effect modification in a study involving several variables, potential confounding or interaction factors.

This type of analysis is considered as an alternative to stratified analysis, since the latter becomes impractical when a large number of variables is to be studied. Considering that it is only possible to estimate the risk for one factor at a time, controlling the set of other variables.

As this report does not have a large number of variables, both approaches were followed and results were compared. As expected, results obtained by the two methodologies were similar. When modeling the logarithm of the odds ratio as a function of exposure and confounding variables, the score statistics derived from the likelihood function are identical to the Mantel-Haenszel test statistics used in stratified analysis. This equivalence is demonstrated by a per-mutational argument which leads to a general likelihood expression in which the exposure variable may be a vector of discrete and/or continuous variables and in which more than two comparison groups may be considered (Day & Byar, 1979).

Stratified Analysis in case-control studies

The essential objective of stratified analysis is to produce an adjusted odds ratio. This is accomplished by determining whether the odds ratios are constant, or homogeneous, over a number of strata.(Iyer, Hosmer, & Lemeshow, 1991).

Any external variable, also a risk factor for the disease under study and also related to the exposure variable, is called a confounding factor or confounding variable as long as it substantially alters the relationship between the disease and the exposure variable. An effect modifying factor or interaction variable is any external variable which, depending on its level, produces significantly different forces of relationship between the disease and the exposure variable.

The stratified analysis is characterized by the fact that the information is divided by strata according to one or more variables, besides the exposure and disease variable.

Consider the following table:

	Exposure		
Disease	Yes (+)	No (-)	**Total**
Cases	a	b	m_1
Controls	c	d	m_0
Total	n_1	n_0	n

Suppose that this table is stratified according to the different i levels, $(1 \leq i \leq k)$ of second risk factor.

Level *i*

	Exposure		
Disease	Yes (+)	No (-)	**Total**
Cases	a_i	b_i	m_{1i}
Controls	c_i	d_i	m_{0i}
Total	n_{1i}	n_{0i}	n_i

If the odds ratios are constant, then a stratified odds ratio estimator such as the Mantel–Haenszel estimator or the weighted logit-based estimator is computed.

The Mantel–Haenszel estimator is a weighted average of the stratum specific odds ratios, $\widehat{OR}_i = (a_i \times d_i)/(b_i \times c_i)$, where a_i, b_i, c_i and d_i are the observed cell frequencies in the 2×2 table for stratum i. The Mantel–Haenszel estimator of the odds ratio is defined in this case as follows:

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i \times d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i \times c_i}{n_i}} \tag{2.17}$$

The logit-based summary estimator of the odds ratio is a weighted average of the stratum specific log-odds ratios where each weight is the inverse of the variance of the stratum specific log-odds ratio,

$$\widehat{OR}_L = \exp\left[\frac{\sum_{i=1}^{k} w_i \ln(\widehat{OR}_i)}{\sum_{i=1}^{k} w_i}\right] \tag{2.18}$$

These estimators provide a correct estimate of the effect of the risk factor only when the odds ratio is constant across the strata (Iyer et al., 1991). Therefore, in the stratified analysis assessing the validity of this assumption is fundamental. Statistical tests of this assumption are based on a comparison of the stratum specific estimates to an overall estimate computed under the assumption that the odds ratio is, in fact, constant (Iyer et al., 1991).

The simplest and most easily computed test of the homogeneity of the odds ratios across strata is based on a weighted sum of the squared deviations of the stratum specific log-odds ratios from their weighted mean. This statistic test, in terms of the current notation, is

$$X_H^2 = \sum_{i=1}^{k} w_i[\ln(\widehat{OR_i}) - \ln(\widehat{OR_L})]^2 \sim \chi^2_{k-1}, \text{under } \mathrm{H}_0.(2.19)$$

Under the null hypothesis that the odds ratio are constant.

Another option is the Breslow and Day test built on the estimated values for α_i $i = 1, \ldots k$

$$\sum_{i}^{k} \left[\frac{(\alpha_i - \hat{\alpha}_i)^2}{\widehat{var}[\hat{\alpha}_i]} \right] \sim \chi^2_{k-1} \tag{2.20}$$

where α_i is calculated through expressions

$$A_i = \widehat{OR}(m_{1i} + n_{1i}) + (m_{0i} - n_{1i}) \tag{2.21}$$

$$B_i = \sqrt{A_i^2 - 4m_{1i}n_{1i}\widehat{OR}(\widehat{OR} - 1)} \tag{2.22}$$

$$\hat{\alpha}_i = \frac{A_i - B_i}{2(\widehat{OR} - 1)} \quad and \quad \widehat{var}[\hat{\alpha}_i] = \left[\frac{1}{\hat{\alpha}_i} + \frac{1}{m_{i1} - \hat{\alpha}_i} + \frac{1}{n_{i1} - \hat{\alpha}_i} + \frac{1}{n_{0i} - m_{i1} + \hat{\alpha}_i} \right]^{-1} \tag{2.23}$$

The test used in *Open.Epi*, the software used for the analyses, is generally referred as the "Breslow-Day test of homogeneity" and is based on a chi square test. To test the interaction of the odds ratio (*OR*), the chi square test is calculated as:

$$\sum_{i=1}^{k} \left[\frac{[\ln(\widehat{OR_i}) - \ln(\widehat{OR_L})]^2}{\widehat{var}\left[\ln(\widehat{OR_i})\right]} \right] \sim \chi^2_{k-1} \tag{2.24}$$

where $\widehat{OR_i} = \frac{a_i d_i}{b_i c_i}$, $w_i = \frac{1}{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}}$, $\widehat{OR_L} = \exp\left(\frac{\sum_{i=1}^{s} w_i \ln(\widehat{OR_i})}{\sum_{i=1}^{k} w_i}\right)$ and the $\widehat{var}[\ln(\widehat{OR_i})] = \frac{1}{w_i}$

The logistic regression model

Logistic regression is one of the most used methods when faced with dichotomous outcomes. It is particularly appropriate for models involving disease state (diseased/healthy) and decision making (yes/no), and therefore is widely used in studies in the health sciences (Bagley, White, & Golomb, 2001).

The binary logistic regression analysis allows the use of a regression model to estimate the probability of a specific event, evaluating the influence of the explanatory variables on the response variable, as well as the effects of their potential interactions. With this methodology, the explanatory variables can be categorical or quantitative (Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997).

The goal of a logistic regression analysis is to find the best-fitting and most parsimonious, yet biologically reasonable, model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables (Lemeshow & David, 2014).

Since the outcome is binary, logistic regression is the appropriate model, and it takes into account that variance changes with the mean. Below a description of the logistic regression model is presented.

In binary logistic regression the dependent variable Y can only take two values: 0 or 1 like this:

$$E(Y) = 0 \times P(Y = 0) + 1 \times P(Y = 1) = P(Y = 1) = p, \tag{2.25}$$

where p is the probability of success event occurrence $(Y = 1)$.

Thus, a reasonable regression model is:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i$$

$$\left(\frac{p}{1-p}\right) = \exp(\beta_0 + \sum_{i=1}^{K} \beta_i x_i)$$

$$p = \frac{\exp(\beta_0 + \sum_{i=1}^{K} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{K} \beta_i x_i)}. \tag{2.26}$$

The logistic regression model (2.26) can be used to describe the relationship between a dichotomous response variable and a set of explanatory variables (i.e age, treatment, gender, etc). This analysis is of great interest in evaluating treatment effects as well as treatment interaction effects with other explanatory variables.

To estimate the model parameters, the maximum likelihood (MLE) method was used. There are other methods for parameter estimation, such as weighted least squares and discriminant function analysis (Hosmer et al., 1997), however these methodologies were not used here.

Considering a sample of dimension n, the likelihood function is given by:

$$\mathcal{L}(\beta \mid x, y, n) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \tag{2.27}$$

where x is the vector $(x_{i1}, x_{i2}, ..., x_{ip})$ and y is the vector $(y_1, y_2, ..., y_n)$.

For a less complex algebraic manipulation, applying the logarithm to the previous expression (2.27) we get:

$$\ln(\mathcal{L}(\beta \mid x, y, n)) = l(\beta \mid x, y, n)) = \sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) =$$

$$= \sum_{i=1}^{n} \left(y_i \ln \frac{p_i}{1 - p_i} + \ln(1 - p_i) \right) \tag{2.28}$$

The next step is to find the $\beta_0, \beta_1, ...\beta_p$ values that maximize the log-likelihood function. For this, the partial derivatives for $\beta_0, \beta_1, ...\beta_p$, are determined.

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial \sum_{i=1}^{n} y_i(\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p) - \ln(1 + \exp(\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))}{\partial \beta_j} \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{\partial l_i}{\partial p_i} \frac{\partial p_i}{\partial \beta_j} = \frac{y_i - p_i}{p_i(1 - p_i)} \frac{\partial p_i}{\partial \beta_j} = x_{ij} \frac{y_i - p_i}{p_i(1 - p_i)} p_i(1 - p_i) = x_{ij}(y_i - p_i), \tag{2.29}$$

and posteriorly equaling zero, we get the following system of equations:

$$\begin{cases} \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_0} = \sum_{i=1}^{n} x_{i0}(y_i - p_i) = 0 \\ \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_1} = \sum_{i=1}^{n} x_{i1}(y_i - p_i) = 0 \\ \qquad\qquad \vdots \\ \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_p} = \sum_{i=1}^{n} x_{ip}(y_i - p_i) = 0, \end{cases}$$

with j=0,1,..., p and $x_{i0} = 1$.

Considering then

$$\mathbf{X}= \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} ; \mathbf{Y}= \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} ; \beta= \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} ; \mathbf{p}= \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} ,$$

this system is denoted by:

$$\mathbf{X}^T(\mathbf{Y\text{-}p}) = 0 \tag{2.30}$$

There is no analytic solution to this system of equations, so it is necessary to use an iterative method to obtain the maximum likelihood estimators (MLE) - the iterative method of least squares, also known as Fisher's score. This method is an adaptation of the Newton-Raphson method in which the Hessian matrix is replaced by the Fisher information (Hosmer et al., 1997).

To infer about the β parameter vector, namely to make hypothesis tests and obtain confidence intervals, it is necessary to know the sample distribution of the MLE.

Restricted cubic splines

In the construction of the binary logistic regression models, the linearity of the logit for continuous variables is required. However, the relationship between the response variable and the independent variable is often non-linear in nature, and the simple inclusion of the independent variable is insufficient to model the relationship.

Splines are a good strategy for the evaluation of non-linear continuous relationships. In this report, we opted for the use of restricted cubic splines.

Restricted cubic splines is a by parts function constituted by polynomials of order 3. These can be restricted or not restricted. In the restricted the tails are modeled by linear functions. In the not restricted, they're not.

A restricted cubic splines regression is defined as a cubic function between adjacent members of a set of fixed knots $t_1 < t_2 < \cdots < t_k$ in the domain of an independent variable X, with a linear function if $x < t_1$ or $x > t_k$, continuous and having first and second continuous derivatives. Normally, 3 to 7 knots are used, so the use of quantiles (tertiles, quartiles, quintiles, etc.) is a good option.

The number of knots is more important than their location, (Frank E. Harrell, 2015) suggests that the number of knots should be selected, based on the sample size. For samples smaller than 100, the use of 4 internal knots is generally efficient, while for larger samples, 5 knots are indicated as a good starting point. From 7 knots on, the analysis can become subjective.

The location of the knots are prespecified based on the quantiles of the continuous variable. This ensures that there are enough observations in each interval to estimate the cubic polynomial (Frank E. Harrell, 2015). Table 2.2 shows suggested locations.

Table 2.2: Locations of knots. (Frank E. Harrell, 2015)

Nr of Knots	**Quantiles**						
3			0.10	0.50	0.90		
4			0.05	0.35	0.65	0.95	
5		0.05	0.275	0.50	0.725	0.95	
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.50	0.6583	0.8167	0.975

In terms of functions of X to be included as independent variables, to express a regression with restricted cubic splines, for example with 5 knots, it implies the addition of three additional variables, denominated in this case by $S_{5,1}, S_{5,2}$ and $S_{5,3}$ and defined by (Korn & Graubard, 2011):

$$S_{5,1} = (X-t_1)^3_+ - \frac{t_5-t_1}{t_5-t_4}(X-t_4)^3_+ + \frac{t_4-t_1}{t_5-t_4}(X-t_5)^3_+$$

$$S_{5,2} = (X-t_2)^3_+ - \frac{t_5-t_2}{t_5-t_4}(X-t_4)^3_+ + \frac{t_4-t_2}{t_5-t_4}(X-t_5)^3_+$$

$$S_{5,3} = (X-t_3)^3_+ - \frac{t_5-t_3}{t_5-t_4}(X-t_4)^3_+ + \frac{t_4-t_3}{t_5-t_4}(X-t_5)^3_+ \tag{2.31}$$

where $(X-t_z)_+$ are equal to $X-t_z$ if $X-t_z>0$, and $(X-t_z)_+$ equal to 0 otherwise, z represent the index of the knots $(z=1,2,3,4,5)$.

Model significance

The decision about which tests to use will depend on whether or not the models are nested.

The likelihood ratio test, is intended to compare two nested models over the same data set. It tests the nullity of a β component subvector r,

$$H_0 : \beta_r = 0 \,\text{vs}\, H_1 : \beta_r \neq 0$$

where the test statistic is given by

$$-2\ln(\mathcal{L}_s/\mathcal{L}_c) = -2(l_s - l_c) \sim \chi^2_{k_c-k_s}, \text{under H}_0 \tag{2.32}$$

where $\mathcal{L}_c$ corresponds to the likelihood of the most general model, with k_c parameters and $\mathcal{L}_s$ corresponds to the likelihood of the (simplest) nested model with k_s parameters. Under the null hypothesis that the constrained model is more appropriate (ie that the $r = k_c - k_s$ additional parameters are null) the statistic of the test has asymptotic distribution of a chi-square with $k_c - k_s$ degrees of freedom.

Another option is the Wald test, that allows to calculate the significance of the coefficients. When applying this test the maximum likelihood estimate is compared, $\hat{\beta}_j$ with an estimate of your standard error. The hypothesis under test is,

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

where the test statistic is given by

$$W = \frac{\hat{\beta}_j}{\sqrt{se(\hat{\beta}_j)}} \tag{2.33}$$

which under H_0 has a gaussian asymptotic distribution, based on estimator properties for β. If the effect of β_j is not significant, then the associated X$_j$ variable is therefore not an important predictor of the response variable.

When dealing with non-nested models, the most commonly used criterion is the Akaike Information Criteria (AIC). By the AIC criterion, the smaller the AIC the better the model. This value is calculated by:

$$AIC = -2\ln\mathcal{L} + 2k, \tag{2.34}$$

Where k represents the number of parameters and $\mathcal{L}$ the likelihood of the model. Note that this criterion penalizes the number of parameters of the model.

Goodness-of-fit

After estimating the model parameters, there are several steps involved in assessing the fit, suitability, and utility of the model.

Examining a model's goodness-of-fit involves determining whether the fitted model's residual variation is small, displays no systematic tendency and follows the variability postulated by the model (Hosmer et al., 1997). Evidence of lack-of-fit may come from a violation of one or more these three characteristics (Hosmer et al., 1997).

The overall quality of the model fit can be verified through the model's AIC value and the application of more general tests that investigate how close the values predicted by the proposed model are to the observed values, sush as the Hosmer-Lemeshow test, often used in Logistic regression, the Pearson chis-square test and the Deviance test. However, some disadvantages are pointed to this classic approach: in Hosmer-Lemeshow tests the value of the statistics depend on the choice of cutpoints that define the groups, in addition, they may have low power for detecting of cutpoints that define the groups (Hosmer et al., 1997).

Considering the disadvantages pointed to the classical methods, new alternatives have emerged, for example Stukel proposed a goodness of fit score test. It is based on the comparison of the logistic model to a more general family of models (Stukel, 1988).

For this test a general logistic model which uses a logit function with two additional parameters was proposed, resulting in the linear logistic model when two parameters are equal to 0.

The proposed score test is a test of:

$H_0 : \alpha_1 = \alpha_2 = 0$, and is denoted by:

$$\widehat{S}_{ST} = \mathbf{s}_s' V_s^{-1} \mathbf{s}_s \tag{2.35}$$

where $\mathbf{s'}_s = \left(\frac{\partial \ln L}{\partial \alpha_1}, \frac{\partial \ln L}{\partial \alpha_2}\right)$, L and V_s is calculated using the Fisher information asymptotic chi-squared distribution with 2 degrees of freedom.

This test was chosen for this study because it was found to be a good goodness of fit test (Hosmer et al., 1997), when compared amongst several competing goodness of fit tests. Hosmer (Hosmer et al., 1997) concluded that $\hat{S}_{ST}$ displayed high power under three types of departures competing amongst several goodness of fit tests.

Residual Analysis

The residuals (e_i) represent the discrepancy of the observed values (y_i) and the predicted values $(\hat{y}_i)$. This is,

$$e_i = y_i - \hat{y}_i, i\text{=}1,...,n. \tag{2.36}$$

Since the response variable is binary, the predicted values vary in the range [0,1] and the model residuals vary in the range $[-1,1]$. When $e_i > 0$, $y_i = 1$ and, for $e_i < 0$, $y_i = 0$. For $e_i = 0$, the adjustment is perfect, that is, y_i=$\hat{y}_i$, the observed values are equal to the predicted values.

In this report, the analysis of residuals aimed to look for potential outliers or influential observations, as these may interfere with the validity and suitability of the model, distorting it.

Studentised Pearson residuals, deviance residuals, leverage values and Cook's distance are key tools in the diagnosis of binary logistic regression. As graphic representations are also a great ally when it comes to diagnosis, for this report residual analysis was mostly done with graphic support.

Pearson residuals (pr_i) are obtained by dividing the ordinary residuals by the estimated standard error of Y_i and defined by,

$$pr_i = \frac{e_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \frac{(Y_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{2.37}$$

This way, the obtained residuals may not have a unit variance, so the ordinary residues are standardized by their estimated standard errors, approximated by $\sqrt{\hat{p}_i(1-\hat{p}_i)(1-h_{ii})}$, coming to the studentised residuals (prs_i). Defined by

$$prs_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-h_{ii})}} = \frac{pr_i}{\sqrt{1-h_{ii}}}, \tag{2.38}$$

where h_{ii} represents the distance between the ith observation in relation to the remaining observations versus its order number.

The leverage values are the values corresponding to the hat matrix values, they are a measure of the importance of an observation in the model fit, ranging from 0 to 1, where the leverage value 1 means that the model is being forced to adjust this observation, which is therefore of great influence. The limit of $2(p+1)/n$ has been proposed, where observations with higher h_{ii} are declared as influential (David A. Belsley, 2004). This measure is therefore important in influential observations.

Another type of residual, very useful to identifying outliers, is the deviance residuals (dr_i). This measures the disagreement between one of the fitted log-likelihood components and the corresponding log-likelihood component obtained if each point were adjusted exactly.

These are defined by

$$dr_i = \text{signal}(Y_i - \hat{p}_i)\left\{-2[Y_i - \ln(\hat{p}_i + (1-Y_i)\ln(1-\hat{p}_i)]\right\}^{\frac{1}{2}}. \tag{2.39}$$

The studentized Pearson residuals, deviance residuals and leverage values are therefore the basis for the diagnosis of binary logistic regression.

A good way to evaluate residuals is its graphical representation.

Pearson's standardized residuals plots as a function of estimated logit values, are one of the most commonly used representations. By observation this type of plots allows to identify potencial outliers, for which special attention should be given. Residuals with an absolute value greater than 2 are considered as potential outliers. However, these plots provides little information about influential outliers, so other measures have to be used. The Cook's distance, is one of the most frequently used.

Cook's distance is a measure of influence based on the estimated value of the regression coefficients. It consists of the standardized difference between $\hat{\beta}$ and $\hat{\beta}_{(-1)}$, which represents the maximum likelihood estimates based on the complete data set and excluding the ith observation, respectively, with subsequent standardization by the covariance matrix of $\hat{\beta}$,

$$\triangle\hat{\beta}_i = (\hat{\beta} - \hat{\beta}_{(-1)})^T(\mathbf{X^T}\widehat{\mathbf{W}}\mathbf{X})\hat{\beta}_i = (\hat{\beta} - \hat{\beta}_{(-1)}) \tag{2.40}$$

Influential plots are widely used in detecting outliers and influential observations. For this report, bubble plots were used. In this plots, the stundentized residuals are represented against the hat values, with the areas of circles representing the observations proportional to the Cook's distance values.

The analysis of the described plots was complemented by a table with the values of residuals, leverage and Cook's distance. These observations and/ or those with higher leverage and Cook's distance values were removed individually or in groups, and the change in model coefficients was analyzed. When the change in model coefficients is high (greater than 10 %), observations are removed individually or in groups and a new model is adjusted.

Predictive capacity of the model

After the adequacy analysis of the model, it is necessary to evaluate its predictive capacity. A widely used methodology is the area under the Receiver Operating Characteristic (ROC) curve.

ROC curve, originating from signal detection theory, shows how the receiver detects the existence of signal in the presence of noise. It plots the probability of detecting true signal (sensitivity) and false signal (1–specificity) for an entire range of possible cutpoints. This measure has become the standard for evaluating a fitted model's ability to assign, in general, higher probabilities of the outcome to the subgroup who develop the outcome (y=1) than it does to the subgroup who do not develop the outcome (y=0). The area under the ROC curve (AUC), which ranges from 0.5 to 1.0, provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not (Iyer et al., 1991).

A plot of sensitivity versus 1–specificity over all possible cutpoints is shown in Figure 2.1. The curve generated by these points is called the ROC curve, and the area under the curve provides a measure that helps us understand if the discrimination is good.

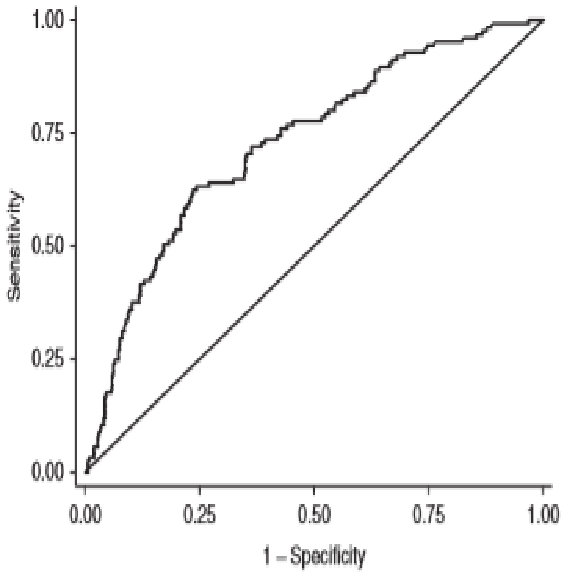


Figure 2.1: Plot of sensitivity vs 1-specificity example.

Normally the discriminatory power of the logistic regression model is used based on the following criterion:

- If AUC=0.5 the model does not discriminate between individuals;
- If 0.6≤ AUC <0.7 the model has limited discrimination;
- If 0.7≤ AUC <0.8 the model presents acceptable discrimination;
- If 0.8≤ AUC <0.9 the model presents good discrimination;
- If AUC ≥0.9 the model presents excellent discrimination.

Model interpretation

Odds Ratio Estimation

An odds ratio (OR) is a measure of association between an exposure and an outcome. The OR represents the odds of the exposure occurring in those with the outcome of interest vs. it occurring in those without the outcome of interest. Odds ratios are most commonly used in case-control studies.

Odds Ratios and Logistic regression

In logistic regression models, the magnitude of the effect of an independent variable $X_i (i = 1, 2, \cdots, p)$ on the response variable can be described by $\exp(\beta_i)$, given that it is shown that

$$\exp(\beta_i) = OR_i, \tag{2.41}$$

where OR_i represents the odds ratio (OR) for X_i adjusted for the other explanatory variables. The OR measures the strength of the association between the dependent variable and any explanatory variable, after discounting the effect of the other variables of the model and verifying their significance.

However, the expression (2.41) is only valid for models without interaction. If the model contains any interaction, the OR of one variable depends on the values of other explanatory variables, it is impossible to describe the effect of a variable through a single value of OR.

If the X_i variable is continuous, supposing that the aim is to express the risk of a certain condition at the x_i' level compared to the x_i'' level for the independent variable by discounting the other variables in the model, the *OR* relating exposure level x_i'' to x_i' is estimated by

$$\widehat{OR} = \exp\left\{\hat{\beta}_i(x_i'' - x_i')\right\}. \tag{2.42}$$

A confidence interval for *OR* can be obtained from the corresponding interval for β_i and is given by

$$\left[\exp\left(\hat{\beta}_i(x_i'' - x_i') - z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\hat{\beta}_i}(x_i'' - x_i')\right), \exp\left(\hat{\beta}_i(x_i'' - x_i') + z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\hat{\beta}_i}(x_i'' - x_i')\right)\right], \tag{2.43}$$

where $z_{(1-\frac{\alpha}{2})}$ is the quantile $(1-\frac{\alpha}{2})$ of the standard normal. In this case, the OR_i relating the exposure level x_i'' to the level x_i' expresses the advantage ratio in favor of the positive value of Y when $X_i = x_i''$ by the advantage when $X_i = x_i'$, keeping the other variables constant.

If independent variable is dichotomous X_i, it should be coded with 1 when present and 0 otherwise.

For the multiple logistic regression model, the OR relates this independent variable with the dependent variable as estimated at

$$\widehat{OR}_i = \exp(\hat{\beta}_i) \tag{2.44}$$

The OR_i represents the comparison between the odds of a positive event occurring on exposure to X_i $(X_i = 1)$ and the odds on non-exposure $(X_i = 0)$, considering that the other variables remain unchanged.

$$\ln(OR)_i = \ln\left(\frac{p}{1-p}\right)\Bigg|_{X_{i=1}} - \ln\left(\frac{p}{1-p}\right)\Bigg|_{X_{i=0}} =$$

$$\beta_0 + \beta_i \times 1 + \sum_{j \neq i} \beta_j x_j - (\beta_0 + \beta_i \times 0 + \sum_{j \neq i} \beta_j x_j) = \beta_i \tag{2.45}$$

A confidence interval for $(1-\alpha)100\%$ to true OR_i is given by

$$\left[\exp\left\{\hat{\beta}_i - z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\hat{\beta}_i}\right\}, \exp\left\{\hat{\beta}_i + z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\hat{\beta}_i}\right\}\right] \tag{2.46}$$

The logistic model can be used to identify the kind of effect that different secondary risk factors have on the link between the major risk factor and the disease. Considering a linear model with a simplified notation:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i.$$

Now considering consider 3 models: a simpler first one with a single independent variable, type (1) (2.47) model, a model with more than one independent variable, type (2) (2.48) model, and finally a type (3) (2.49) model that contains an interaction factor:

Type (1) model

$$Y = \beta_0 + \beta X_1 \tag{2.47}$$

Type (2) model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{2.48}$$

Type (3) model-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{2.49}$$

Based on the presented models, in order to verify whether X_2 is an effect modifier or an confounding, in the association between Y and X_1, we must verify whether:

1. If in the type (3) model (2.49) the value of β_3, coefficient of interaction, is close to zero and if β_3 is not significantly nonzero, variable X_2 is not an effect modifying factor, otherwise it will be.

2. If factor X_2 is not an effect modifying factor, to see if it is a confounding factor just check if β_1 is different in type model (2) (2.48) and type model (1) (2.47). If it's the same, it's not a confounding factor, if it's not the same, it's a confounding factor.

2.1.3 Multinomial Logistic Regression

The multinomial logistic regression model, as explained by the IAUL & Department of Statistics (2002) consists in adapting the logistic regression model to a nominal response variable with more than 2 levels.

The objective is to model the choice odds of each level of the response variable as a function of the covariates and to express the results through the odds ratio, according to the choice of different levels.

The first step is to select one category as the reference category. The non-references categories are defined as:

$$\ln(\pi_j) = \ln\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x_{1j} + ... + \beta_{pj}x_{pj}, \quad \text{for } j = 2,...,J. \tag{2.50}$$

This represents $J-1$ logit equations that are used simultaneously to obtain estimates of the parameters β_j. The parameters estimates can then be used to obtain estimates for the category probabilities $\hat{\pi}_1,...,\hat{\pi}_J$, subject to the constraint that $\hat{\pi}_1 + \hat{\pi}_2 + ... + \hat{\pi}_J = 1$. This is done by rewriting the above model, so that the π_is are the responses. After estimating the probabilities, the adjusted values for the model can be calculated by multiplying each of the estimated probabilities by the total number of observations, n.

Odds ratio and model interpretation

Odds ratios generally offer much easier interpretation than directly interpreting model parameters. In order to explain the use of odds ratios, we will consider a simple model involving a response variable with J categories and a binary explanatory variable x . The explanatory variable x denotes the presence ($x = 1$) or absence ($x = 0$) of some 'exposure' factors. The odds ratio for 'exposure' for response $j(j = 2...,J)$ relative to the reference category $j = 1$ is defined as:

$$\theta_j = \frac{\frac{\pi_{j|_{x_i=1}}}{\pi_{j|_{x_i=0}}}}{\frac{\pi_{1|_{x_i=1}}}{\pi_{1|_{x_i=0}}}} \tag{2.51}$$

where π_{jp} and π_{ja} denote the probabilities of response category j ($j = 1,...,J$) for the cases where the exposure is present or absent, respectively. The model for our example can be specified as:

$$\ln\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x, \quad j = 2,...,J. \tag{2.52}$$

The logarithm of the odds ratio for the above model can be written as:

$$\ln(\theta_j) = \ln\left(\frac{\pi_{jp}}{\pi_{1p}}\right) - \ln\left(\frac{\pi_{ja}}{\pi_{1a}}\right) = \beta_{1j}. \tag{2.53}$$

Chapter 3

The Data

This chapter, presents the data set, its pre-processing and the most relevant variables used and created for this report.

3.1 Data Colection and Dataset

The dataset comprises 2250 participants, 1326 ALS patients and 924 controls from various European nationalities who answered a standardized questionnaire, created in 2015 as part of the OnWebDuals project, with the aim of collecting ALS patient data throughout European locations, to build an ALS domain ontology, and implement it in a large pan-European web-database. This was the basis of the information worked on all patients and controls.

The questionnaire used in the OnWebDuals project was validated by an international panel of neurologists in Europe, Japan, Australia, South Africa, South America and the USA and applied by experienced neurologists during the anamnesis. The project was approved by the relevant Ethics Commission. All patients and controls signed a written informed consent before inclusion in the study (DeCarvalho et al., 2017).

The questionnaire features 160 items regarding patient demographic and clinical information and ALS risk factors. Questions are dived in 3 categories: 1 – essential; 2 – important; or 3 – not very relevant. Participants were free to add a few additional items they thought could be relevant (DeCarvalho et al., 2017).

Dataset

The data arrived in 2 different datasets, one for patients and one for controls. These datasets are CSV files that compiles patients and controls answers to the questionnaire. The patients dataset has 1326 rows, representing each of the patients trough a unique alphanumeric code, and 193 columns, based on a question or part of a question from the survey. Same for the control data set with 924 rows, representing each of the controls trough a unique alphanumeric code, and 169 columns, also based on a question.

The two datasets are organized in the same way, with a different number of columns related to the exclusive clinical information of the patients. For easier management and further analysis, the databases were aggregated into one. Patients and Controls are distinguished by the introduction of a new dichotomic variable, `Type` where 0 represents Controls and 1 represents Patients.

Before starting data analysis, inconsistencies in the variables were checked, particularly in the candidate variables for analysis. Some inconsistencies were related with differences between the expected answers and the given answers. When the expected answer was “Yes” or “No”, categorized in the database with the variables “No”(0),“Yes”(1), instead of 0’s and 1’s, some “Positive” or “Negative” were identified. A very frequent variable in the database is the date, like consultation date, date of first symptoms; of smoking start or stop, among others. To avoid inconsistencies in this kind of variables, all dates must be in the same format like “YYYY/MM/DD” or “DD/MM/YYYY”, for example. It is also important to assure that if the unit of the variable is the year, it is expected that it is not less than 1900.

The presence of empty cells or cells with “NR” (Not relevant), “NA” (Not available) or “NF” (Not viable) were also verified to identify missing entries. For controls, all patient-exclusive variables were also empty.

Controls ‘ANTC-0044’, ‘ANTC-0048’ and ‘ANTC-0057’ have negative ages. The age corresponds to the age at the consultation date. For these participants, the consultation date is before the date of birth. As it was not possible to validate this information, these controls were removed from the database.

From the variables present in the original database, new variables were generated, as described below.

3.2 Variable Coding

Sociodemographic variables

The selected sociodemographic variables provide relevant information for the characterization of the participants and their lifestyle. The importance of variables such as age, gender and smoking habits are usually referred as risk factors for ALS.

Age of the participants at the time they answered the survey, the identification of gender and smoking habits were selected for this report. The `Smoking` variable with 3 levels (0-‘Non smoking’ 1-‘Smoking’ and 2-‘Ex-smoking’) as worked was derived by grouping the `Smoking` variable (0 - ‘Non smoking’, 1 -‘smoking’), which tells if the person smokes or does not smoke and the variable `Stopped smoking` (0 - ‘No’, 1 - ‘Yes’) when knowing that the person has quit smoking, we move it to the ex-smoking category. For participants with sufficient information, the exposure load `TE` was calculated in pack-years. Pack-years, calculated by multiplying average packs smoked per day by the duration of smoking, in years, using the following expression (Nance et al., 2017):

$$\text{Number of pack-years (TE)} = (\text{packs smoked per day}) \times (\text{years as smoker})$$

The variable , with the highest number of missing cases, was TE variable, with 234 missings.

Table 3.1: Characterization of sociodemographic variables

	Variable	Meaning	Codification
Age	`Age`	Participant's age at the date of the questionnaire	
Gender	`Gender`	Participant gender	0-Female, 1-Male
Smoking	`Smoking`	Participant Smoking habits	0-Non Smoking, 1-Smoking, 2-Ex-Smoking
Tobacco exposure	`TE`	Quantification of cigarette smoking in pack-years	

Still regarding the habit of smoking, it seems interesting to see how it is distributed at the time of diagnosis, before and after it. This analysis was made trough the graphs present in fig. 3.1.

In this fig 3.1 each line represents the number of years of smoking for each participant. For participants who stopped smoking, the lower limit corresponds to the date they started smoking and the upper limit to the date they stopped smoking, for participants who have not stopped smoking at the date of the consultation, the upper limit corresponds to the date of the consultation. The vertical line at 0 corresponds to the date of diagnosis. Negative values represent years before diagnosis.

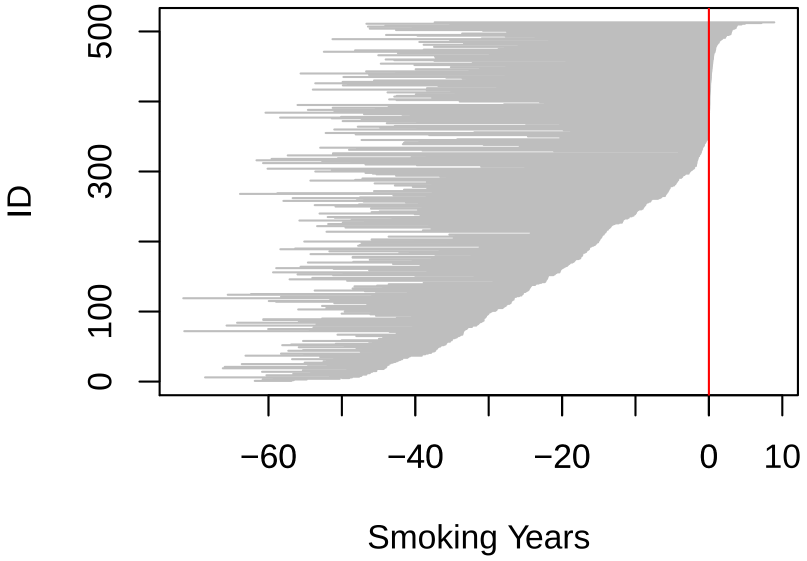


Figure 3.1: Distribution of smoking habit at diagnosis, before and after, 0 represents the diagnosis date.

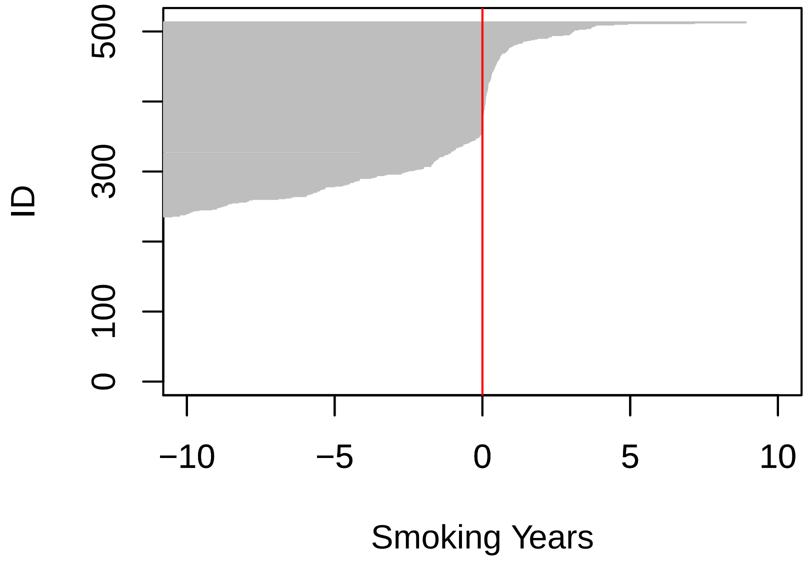


Figure 3.2: Distribution of smoking habit at diagnosis, restrict the interval to 10 years before and 10 years after the diagnosis.

Based on fig. 3.1, although there are patients who continue to smoke after diagnosis and participants who stopped smoking long before the diagnosis date, it seems that many patients stop smoking near the diagnosis date or at the diagnosis date.

When restricting the interval to 10 years before and 10 years after the diagnosis (fig. 3.2), this relationship is more noticeable. These results meet our expectations, as it is understood that with the first symptoms patients start to not feel very well, and decide to quit smoking.

Clinical Variables

As there are a large number of clinical variables in the database, only those of interest for the present report were selected. These variables are only for the group of patients.

Regarding genetic information, the variables that evaluate the presence or absence of genes with the highest ALS association were selected and represented by the variables `C9orf72` and `SOD1` both dichotomic (0-'No', 1-'Yes') where 0 represents the absence of mutation and 1 the presence of mutation. Regarding the start type, we have the `Onset` variable (1-'Limbs', 2-'Bulbar') resulting from grouping the `Limbs Onset` and `Bulbar Onset` variables, although there are other onset types, these are the most common ones.

The `Age.1st.Sym` and `Diag.Age` variables correspond to the conversion into years / ages of the original `Date of 1st Symptoms` and `Date of Diagnosis` variables, obtained by the difference between these and the patient's date of birth `Date of birth`.

Still within the clinical variables, in relation to disease progression levels, there's the `ALSFRS.R.T` variable, corresponding to the score obtained in the questionnaire to assess the patients' functional level and the variable `ALSFRS.R`, with the values obtained from the expression given in (1.1) for each patient. Regarding the level of progression, there's `PG` variable, a three-level categorical variable (0-'Slow', 1-'Neutral' and 2-'Fast') based on the values of variable `ALSFRS.R`, patients with values between 0 and 0.5 are part of the slow progression group, between 0.5 and 1.5 neutral progression group and greater than 1.5 fast progression group.

Table 3.2: Characterization of Clinical variables

	Variable	Meaning	Codification
Abnormal C9orf72 repeat-expansion	`C9`	Patients mutation	0-No, 1-Yes
SOD1 mutation	`SOD1`	Patients mutation	0-No, 1-Yes
Onset	`Onset`	Type of disease onset	1-Limbs, 2-Bulbar
Age of 1st manifestation	`Age.1st.Sym`	Age of patient on date of first symptom	-
Diagnosis age	`Diag.Age`	Age of patient on date of diagnostic	
Diagnostic delay	`Diagnostic.delay`	Diagnostic delay in months	
ALS Function Rating Scale score	`ALSFRS.R.T`	Score obtained in the questionnaire to assess the patients' functional level	
ALS Function Rating Rate of decay	`ALSFRS.R`	Progression Rate obtained based in expression 1.1	
Progression Group	`PG`	Represents disease progression levels	Slow, Neutral, Fast

Sports Variables

Although there are 21 different modalities in this database, the number of participants in each one of them is very small. The most represented modality is football, with 183 participants.

Initially grouping the 21 modalities into high impact and moderate or mild impact modalities was considered a good option. However, this option became quite subjective because most of the identified sport division or classification are based on levels of competitive play (professional or non-professional), intensity (high/low) or joint impact. Since dealing with a neurodegenerative disease, the focus was only on the contact modalities, because although the relationship between physical activity and ALS is still somewhat contradictory, it is thought that contact modalities at the highest competitive levels, that combine intense physical activity and increased risk of repetitive head and cervical spine trauma are a risk factor for neurodegenerative diseases Blecher et al. (2019). Based on the work of (Blecher et al., 2019), four modalities were selected: boxing, hockey, football and rugby.

The 'problem' of a small number of participants remains when considering the modalities of contact. Instead of working the variables for each of the modalities thought, their information was condensed into the `Contact` variable. With greater interest in the highest intensity levels, the `Contact` variable is a two-level categorical variable ('Yes' and 'No'), where 'Yes' represents moderate to high intensity contact practitioners corresponding to the `Vigorous` and `Moderate` in the database, and 'No' represents low intensity contact practitioners, `Mild`, participants who do not practice any modality, and practitioners of any of the other modalities regardless of the intensity with which they are practiced.

An exclusive variable was created for football players, as it is an important factor in the goals of the study and is also the most widely practiced and of global interest. `Football` variable is made up of 3 levels 'No practice', represents participants who do not practice any activity or practice only one of the modalities present in the database besides football, 'Mild' representing a practitioners in a more recreational aspect without great intensity and finally 'Intense' representing professional football players or with a high level of practice.

Note that the practice of these modalities could be for different durations and at different times in the past, but this information is not detailed enough for analyses. Thus, we consider together all those who indicated in the questionnaire that they practiced these modalities at some point in their lives

Table 3.3: Characterization of Sports variables

	Variable	Meaning	Codification
Contact sports	`Contact`	Participants practicing football, hockey, boxing, rugby with different intensities	0-No, 1-Yes
Football	`Football`	Level of intensity at which participants play football	0-No practice, 1-Mild activity, 2-Intense activity

Chapter 4

Results

This chapter presents the results. First, a brief view of how the data behaves is given by the presentation of the results of the exploratory analysis - number of participants, number of elements by gender and average age, among others. Stratified analysis in case control studies and binary logistic regression models were used, to evaluate the relationship between the practice of contact sports and ALS and the existence of variables that can interfere in this relation namely age, gender and smoking habits.

To answer the second objective - to assess the relationship between age of early diagnosis and the practice of high-intensity football, ANOVA was used. Linear regression and multinomial logistic regression models were used, to evaluate the variables associated with delayed diagnosis and different levels of disease progression.

4.1 Exploratory Analysis

The sample consists of 2247 individuals, of which 1326 are part of the patient group (768 men and 558 women) and 921 of the controls (430 men and 491 women). For some variables discussed below, the total may be less than the 2247 due to missing values.

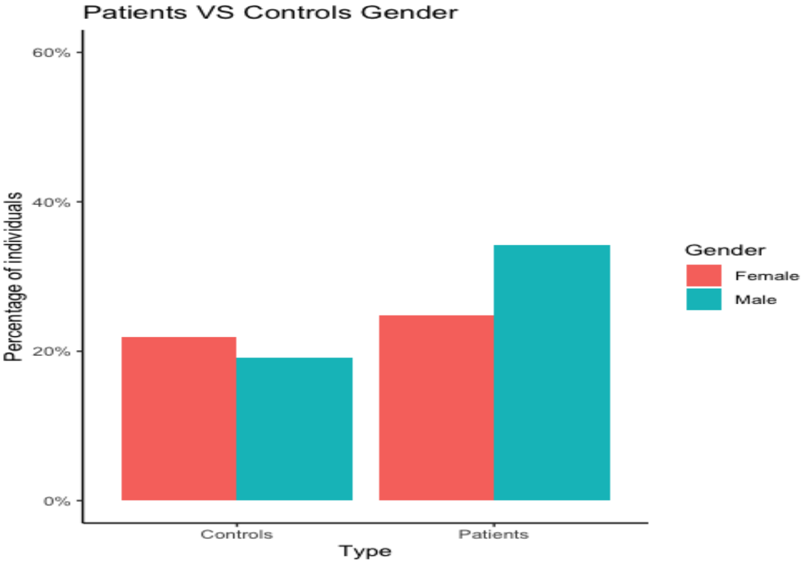


Figure 4.1: Distributions individuals by gender. In the patients, 42.1% are women and 57.9% are men, whereas in the controls, we have 53.3% women and 46.7% men.

Considering the entire population, the average age at the date of consultation when individuals were questioned is 59.46 years.

In women the average age for patients is 63.5 years, with a standard deviation of 12.8 years, the median age is 65.3 years, the oldest patient who answered the questionnaire is 89.8 years and the youngest 20.7 years. The average and median are very close as seen in fig.4.2. In this group, there are 7 outliers. In the control group the average age in women is 55.1 years, with a standard deviation of 15.2 years, the average age is 56 years, with a maximum of 91.4 years and a minimum of 19.1 years. As in the patient group, the distribution is approximately symmetrical.

In men the average age for patients is 60.7 years with a standard deviation of 12.8 years, the median age is 61.7 years, the maximum is 89.4 years and the minimum 17 years. The age distribution in the group is approximately symmetrical. There are 3 outliers for whom the average age is lower than for the remaining. In controls the average age for men is 57 years with a standard deviation of 15 years, the median age is 59.3 years, the maximum is 86.6 years and the minimum 12. The average is lower than the median, although the difference is not very large. In this group, we have a slight asymmetry on the left and an outlier candidate.

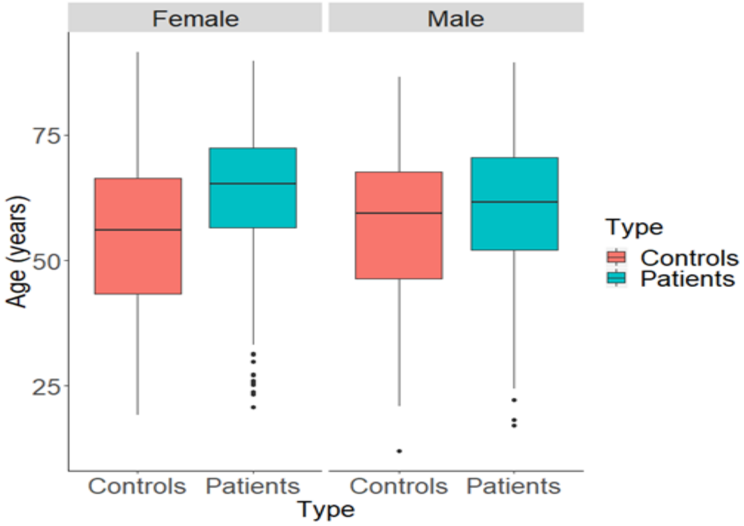


Figure 4.2: Age in years at the date of consultation when individuals are questioned for patients and controls by gender.

Regarding smoking habits, there are 1274 non-smokers (738 Patients and 536 Controls), 322 smokers (171 Patients and 151 Controls) and 634 (402 Patients and 232 Controls) former smokers.

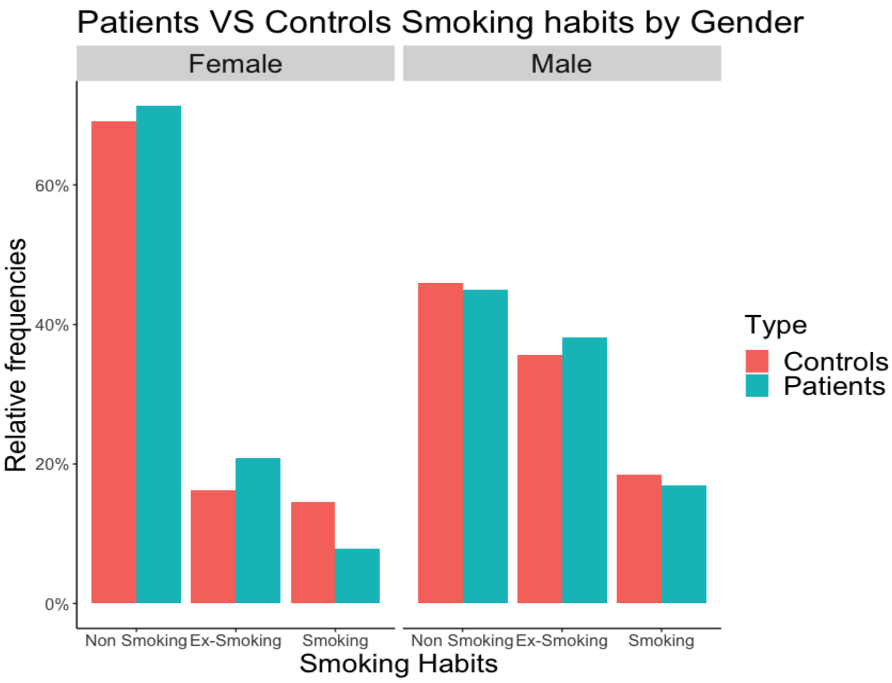


Figure 4.3: Distributions individuals by smoking habits, type and gender. Among women, 70.1% never smoked in the group of patients, 20.6% were former smokers and 7.7% smoked; in controls 69% do not smoke, 16% quit and 14.7% smoke. In male patients 44.5% do not smoke, 37.4% are former smokers and 16.7% smoke; in controls 45.8% don't smoke, 35.6% are former smokers and 18.4% are smokers.

For smoking habits and contact sports, in patients practitioners, there are 56 non-smokers, 36 smokers and 60 ex-smokers. In the controls, there are 35 non-smokers, 12 smokers and 22 ex-smokers. In non-practitioners, there are 675 non-smokers, 134 smokers and 342 ex-smokers in the group of patients, while for the controls there are 498 non-smokers, 138 smokers and 210 ex-smokers. Considering exclusively the male population, in smokers, including smokers and ex-smokers (as they were smokers before), there are 78 participating patients, 19 controls and 245 non-practitioners patients in addition to 157 controls. In non-smokers there are 50 practitioners in patients and 25 in controls, compared to non-practitioners, there are 284 patients and 172 controls.

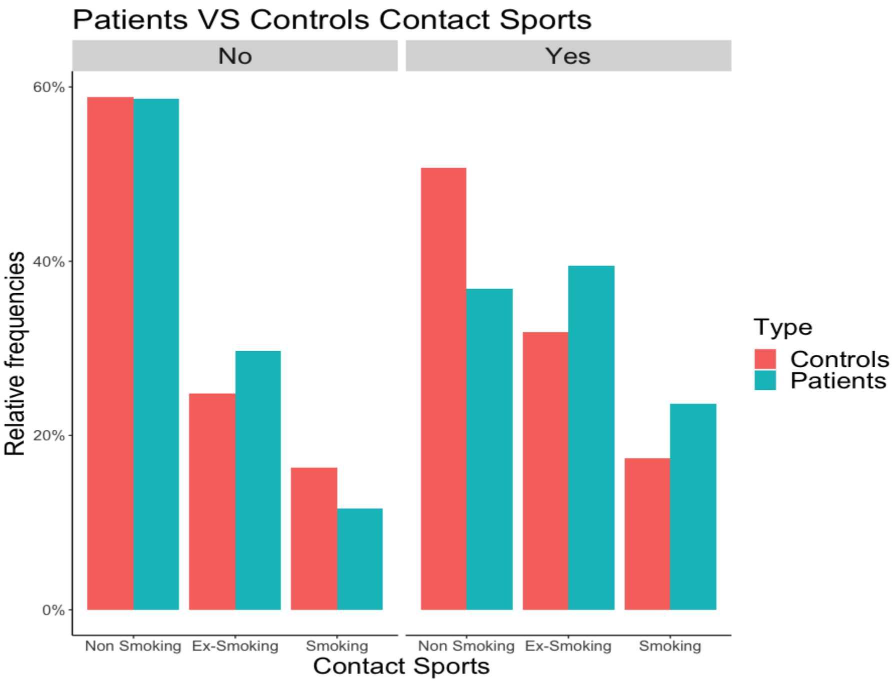


Figure 4.4: Distributions individuals by smoking habits, type and contact sports. Among contact sports practioners 36.84% never smoked in the group of patients, 39.47% were former smokers and17.39% smoked; in controls 50.72% do not smoke, 31.88% quit and 17.39% smoke. In non-practioners 58.64% do not smoke, 29.71% are former smokers and 11.64% smoke; in controls58.87% don't smoke, 24.82% are former smokers and 16.31% are smokers.

Considering the entire population, the average TE is 9.67 packs-year. For non-smokers TE is zero. In the group of patients for former smokers the average is 26.29 packs-year, with a standard deviation of 26.26 packs-year, the median load was 20 packs-year, with the maximum load of 174 packs-years and a minimum of 0.05 packs-years, the average TE is higher than the median TE. There is a slight positive asymmetry and it is also important to highlight the presence of 9 potential outliers. In smokers the average load is 29.12 packs-year, with a standard deviation of 29.12, the median is 19.1 packs-year. Finally, the maximum load recorded is 116 packs-year and the minimum 1.65 packs-year. As for ex-smokers, there is a slight positive asymmetry in the distribution of this group, where there are two potential outliers.

In the control group, ex-smokers have an average TE of 21.05 packs-years, with a standard deviation of 19.64 packs-year, median 15 packs-year, maximum 123 packs-year and minimum 0.1 packs-year. The average TE is higher than the median TE so we are dealing with a case of positive asymmetry and we have 6 candidates for outliers. Smokers have an average TE is 31.57 packs-year, with a standard deviation of 29 packs-year, the median load was 22.5 packs-year, maximum 100 packs-year and minimum 0.1 packs-years. The average and median are close, so the distribution is approximately symmetrical with 5 potential outliers observed.

4. RESULTS

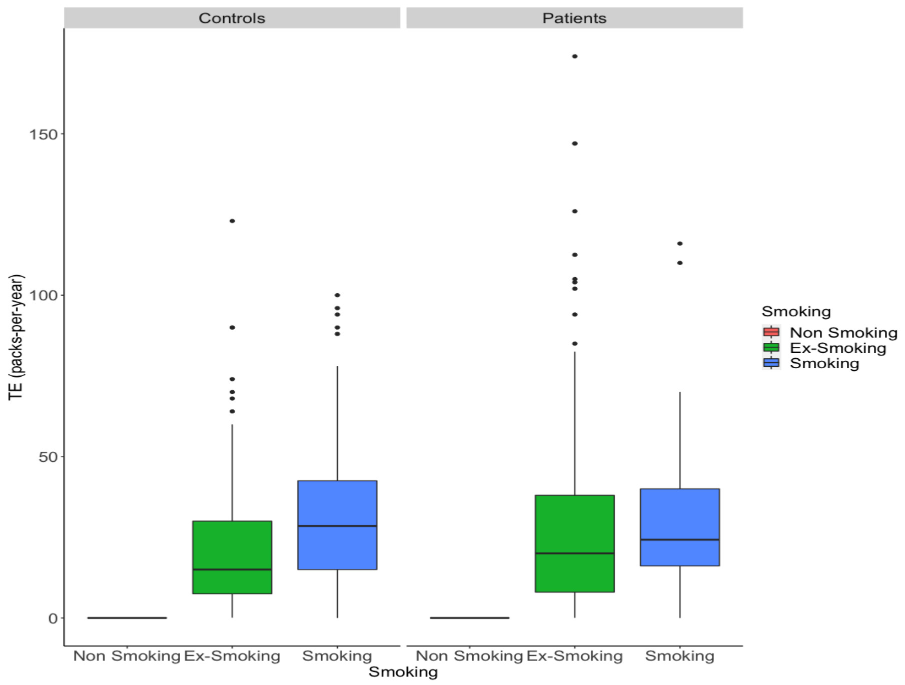


Figure 4.5: Boxplot for TE (packs-years) for different smoke categories in patients and controls.

Limb onset was found in 925 patients (585 men and 340 women), 335 had bulbar onset (145 men and 190 women).

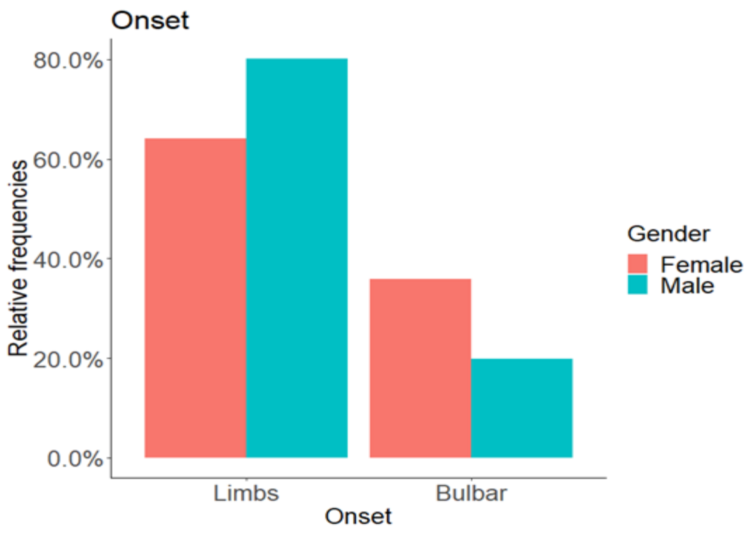


Figure 4.6: Distribution of individuals resulting from the crossing of the gender with the onset of the disease. In the female, 64.1% had onset in the limbs and 35.8% the onset was bulbar; in the men 80.2% had initial limbs in relation to 19.8% with have bulbar onset.

Considering the entire population, the average age of the first manifestation is 59.5 years.

In women with onset of limbs, the average age of the first symptom is 58.4 years, with a standard deviation of 14.2 years and a median of 59.7 years, the maximum was 88.4-year-old and the minimum 16.1-year-old. The mean and median are very close, and the age distribution of the diagnosis for women with initial limbs is approximately symmetrical, this group has 6 possible outliers. In women with bulbar onset, the average age of the first symptoms was 65.7 years, with a standard deviation of 11.1 years, the maximum was 89.2 years and the minimum was 24.6 years old. The mean and median are very close, in this group there are 5 outliers candidates.

In men with affected limbs, the average age at first symptom was 56.6 years, with a standard deviation of 13.7 years, the median of 57.1 years, the maximum of 88.4 years and the minimum of 8.9 years. The distribution in this group is symmetrical, with 5 outliers candidates. For those with bulbar onset, the average is 62.8 years, with a standard deviation of 10.8 years, the maximum is 82.9 years and the minimum 29.1. As for limbs onset, the distribution is symmetrical, and in this group there are only 1 outlier candidate.

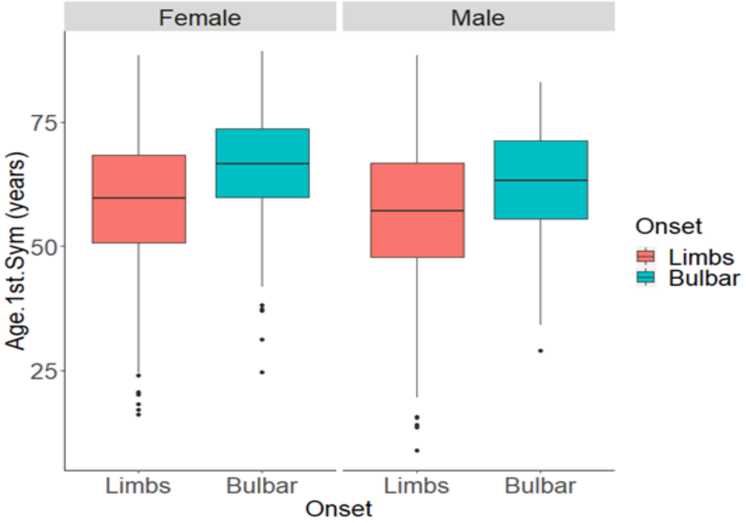


Figure 4.7: Boxplot for age of first symptoms for onset type grouped by gender

The diagnostic delay is higher for the patients with limb onset compared to patients with the bulbar onset. The average diagnostic delay for patients with limbs onset is 1.7 years, with a standard deviation of 2.2 and the median is 1, there is a slight positive asymmetry, the maximum is 25.1 years and the minimum of 0. In patients with bulbar onset the average diagnostic delay is 0.9 years with a standard deviation of 0.8, the median is 0.7. Although average and median are close, there is a slight positive asymmetry as the maximum is 5.3 years and the minimum 0.1 years. For both limb onset and bulbar onset, there are possible outliers (more in the limbs onset).

To better see the differences between the two groups, the elements with the diagnostic delay greater than 5, were removed in fig.4.8b). Therefore, as already mentioned, the diagnostic delay is higher for the patients with limb onset compared to patients with the bulbar onset.

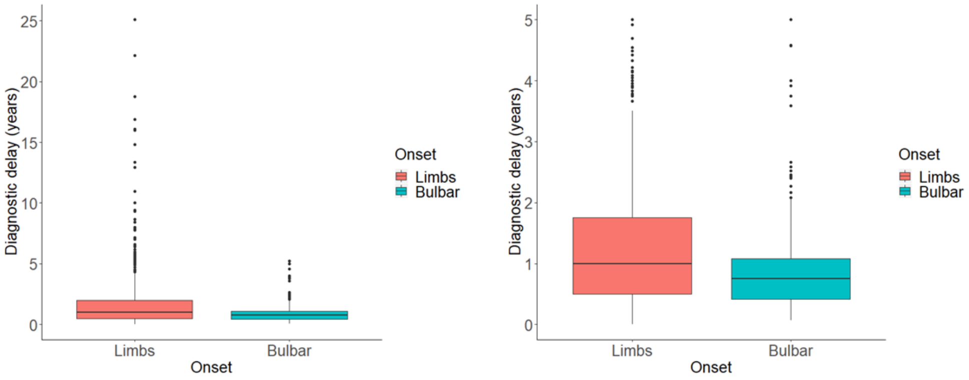


Figure 4.8: **a)** Boxplot for diagnostic delay for onset type. **b)** Boxplot for diagnostic delay $<= 5$ for onset type.

The average ALS Functional Rating Scale-Total is 35.63 and the ALS Functional Rating Scale-Rate of Decay is at 0.83. Concerning the progression groups, there is a first group with slow progression composed of 615 elements, a neutral progression group with 498 elements and a fast progression group with 175 elements.

For the fast progression group the average ALSFRS.R is 2.82 with a standard deviation of 1.55, the median is 2.19, with a maximum of 11.91 and a minimum of 1.7. Although the average and the median are very close there is a slight positive asymmetry, 9 possible outliers are visible.

In the neutral progression group the average is 0.84, with a standard deviation of 0.25, the median is 0.79, as in the previous and median group are very close the distribution is approximately symmetrical; the highest value is 1.49 and the lowest 0.25.

Finally, for the slowly progressing group, the group with the highest expression, the average is 0.277, with a standard deviation of 0.14 the median is 0.5, the maximum is 0.49 and the minimum 0.

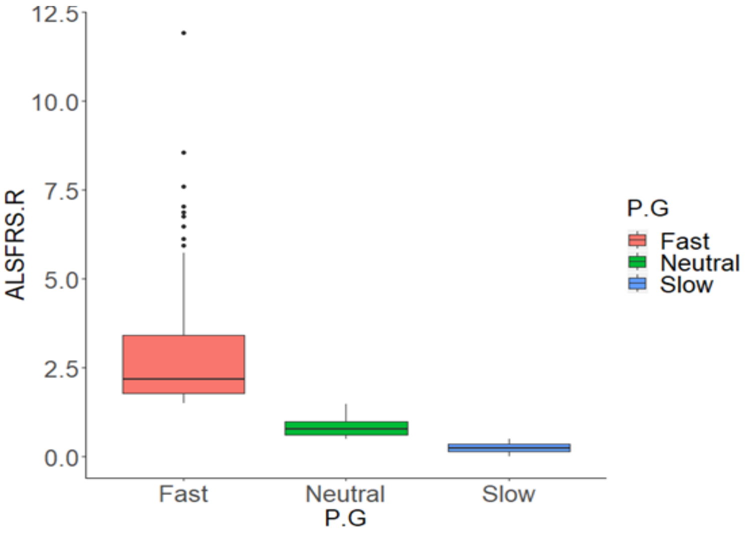


Figure 4.9: Boxplot for ALS Functional Rating Scale-Rate of Decay for Progression Group

Considering the female population the average ALSFRS.R is 0.91 with a standard deviation of 1.05, the median is 0.6, there is a slight positive asymmetry, the maximum is 11.91 and minimum is 0. For men, the average ALSFRS.R is 0.77 with a standard deviation of 1.01, the median is 0.48, the maximum is 8.55 and the minimum is 0.

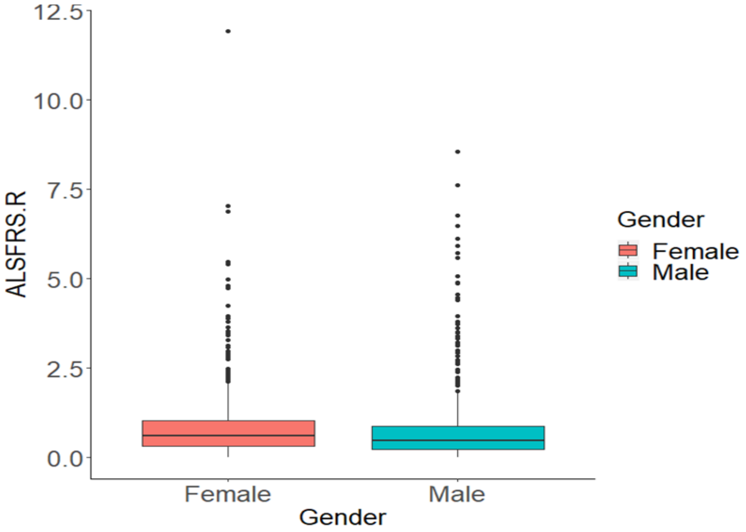


Figure 4.10: Boxplot for ALS Functional Rating Scale-Rate of Decay by Gender

Regarding physical activity, 1016 individuals have regular physical activity, of which 221 practice contact sports intensely. Football in isolation was also evaluated. Of the 183 football players, 10 are practitioners with a low intensity level and 173 with moderate to high intensity.

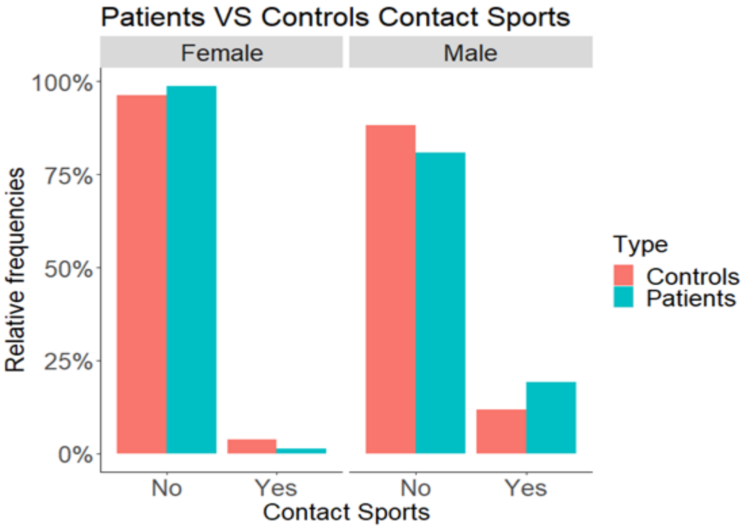


Figure 4.11: Distribution of individuals by the practice of contact sports, type and gender. In female patients, only 1.3% practice contact sports, while in controls 3.7% practice this activity. In men, 19.2 % of patients play contact sports and 11.9% of controls.

4.2 Evaluate the relationship between the practice of contact sports and ALS and the existence of variables that can interfere in this relation: age, gender and smoking habits

Entering in our first objective - to evaluate the relationship between the practice of contact sports and the disease and the existence of variables that can interfere in this relationship such as age, gender and smoking habits - stratified analysis in case-control studies and binary logistic regression were used. First, the results obtained through the stratified analysis are presented, followed by binary logistic regression.

4. RESULTS

Stratified Analysis in case-control studies

For the stratified analysis, the first step is the construction of the strata. In this case, the exposure was the practice of contact sport and the aim was to evaluate the effect of age and gender on the relation between contact sports and ALS. Age was converted into categorical, consisting of 2 levels < 55 and 55^{+}.

Table 4.1: Unstratified (Crude) Values

Contact

Disease	Yes (+)	No (-)	**Total**
Cases	152	1153	1305
Controls	69	846	915
Total	221	1999	2220

Table 4.2: **Stratum 1** Women < 55

Contact

Disease	Yes (+)	No (-)	**Total**
Cases	4	115	119
Controls	7	220	227
Total	11	335	346

Table 4.3: **Stratum 2** Women 55^{+}

Contact

Disease	Yes (+)	No (-)	**Total**
Cases	3	430	433
Controls	11	248	259
Total	14	678	692

Table 4.4: **Stratum 3** Men < 55

Contact

Disease	Yes (+)	No (-)	**Total**
Cases	50	148	198
Controls	23	187	210
Total	73	335	408

Table 4.5: **Stratum 4** Men 55^{+}

Contact

Disease	Yes (+)	No (-)	**Total**
Cases	95	421	516
Controls	28	230	258
Total	123	651	774

Table 4.6: Results from OpenEpi, Version 3, open source calculator-TwobyTwo;

Odds-Based Estimates and Confidence Limits

	Point Estimates		**Confidence Limits**	
Stratum	**Type**	**Value**	**Lower, Upper**	**Type**
1 Women < 55	CMLE OR*	1.093	0.275, 3.851	Mid-P Exact
			0.23, 4.404	Fisher Exact
	OR	1.093	0.314, 3.811	Taylor series
2 Women 55^{+}	CMLE OR*	0.158	0.035, 0.539	Mid-P Exact
			0.028, 0.605	Fisher Exact
	OR	0.157	0.044, 0.569	Taylor series
3 Men < 55	CMLE OR*	1.718	1.008, 2.987	Mid-P Exact
			0.977, 3.094	Fisher Exact
	OR	1.721	1.004, 2.949	Taylor series
4 Men 55^{+}	CMLE OR*	1.852	1.189, 2.947	Mid-P Exact
			1.164, 3.025	Fisher Exact
	OR	1.854	1.181, 2.91	Taylor series
Crude	CMLE OR*	1.616	1.203, 2.186	Mid-P Exact
			1.19, 2.211	Fisher Exact
	OR	1.616	1.2, 2.177	Taylor series
Adjusted	CMLE OR*	1.433	1.054, 1.961	Mid-P Exact
			1.043, 1.985	Fisher Exact
	Directly Adjusted OR	1.492	1.08, 2.06	Taylor series
	Mantel-Haenzel OR	1.431	1.051, 1.985	Robins,Greenland,Breslow
	Breslow-Day test for interaction of Odds Ratio over strata:			
	chi-square=	13.15	**p=0.004**	
	p is less than 0.05, suggesting that OR differs among strata(interaction).			

Note:Conditional Maximum Likelihood Estimator (CMLE)

Considering the obtained results, presented in table 4.6, the Breslow and Day test rejects stratum homogeneity, p-value (0.004). Thus, there is an interaction, that is, an effect change in the relationship between the practice of contact modalities and the disease. Based exclusively on these results, we do not know whether this change is due to the difference in age or gender.

However, looking at the OR values present in table 4.6, we see that strata 3 and 4 are not very different, which leads to that if we consider only the male population, possibly there would be no interaction with age. Therefore gender may be the potential effect modifier. However, it is important to note that the number of women practicing contact modalities is very small, which may explain this difference between genders

Although these measures cannot be used, because the strata's homogeneity is rejected, through the table 4.6, the odds ratio of each stratum and the overall odds ratio can also be seen. In women <55 years old, the odds of developing the disease in contact sports practitioners is 1.093 times higher than the odds of disease in women of the same age group who do not practice contact sports. In men in the same class age (<55), the odds of developing disease in practitioners is 1.72 times higher comparing with non-practitioners. In the second age range 55^+, the odds of developing the disease in women practitioners is 0.157 times lower than in non-practitioners, in men is 1.85 times higher.

Based on the global table 4.1, considering only the practice of contact activity, the odds of developing the disease in practitioners is 1.62 times higher than non-practitioners.

Logistic Regression

In order to compare the results obtained by stratified analysis with logistic regression, the fitted model (4.1) includes the same variables used in the stratified analysis.

Model I:

$$\ln\left(\frac{p}{1-p}\right) = -0.649 + 0.089 ContactYes_i + 0.883 GenderMale_i + 1.199 Age_i 55^+ + 0.454 ContactYes_i \times GenderMale_i - 1.939 ContactYes_i \times Age_i 55^+ -0.828 GenderMale_i \times Age_i 55^+ + 2.013 ContactYes_i \times GenderMale_i \times Age_i 55^+ \tag{4.1}$$

Table 4.7 presents the estimates of the Model I coefficients, as well as the p-value for the Wald test, the *OR* estimate and the corresponding 95% confidence interval estimate.

Table 4.7: Parameter estimates, Wald test-p-values, odd ratio (*OR*) and respective 95% confidence intervals of Model I (Patients Vs Controls)

Variable		β	p-value	*OR*	CI *OR* (95%)
Intercept		-0.649	$<<0.001^{***}$	0.523	(0.416; 0.653)
Contact					
	No	Ref			
	Yes	0.089	0.888	1.093	(0.282; 3.697)

4.2 Evaluate the relationship between the practice of contact sports and ALS and the existence of variables that can interfere in this relation: age, gender and smoking habits

Gender					
	Female	Ref			
	Male	0.883	<<0.001***	2.417	(1.772; 3.309)
Age					
	< 55	Ref			
	55^+	1.199	<<0.001***	3.317	(2.526; 4.374)
ContactYes×GenderMale		0.454	0.513	1.574	(0.416; 6.724)
ContactYes×Age55$^+$		-1.939	0.034*	0.144	(0.002; 0.861)
GenderMale×Age55$^+$		-0.828	<<0.001***	0.437	(0.297; 0.64)
ContactYes × GenderMale×Age55$^+$		2.013	0.043**	7.487	(1.097; 55.64)

What can we get from our model:

1) The ContactYes interaction with Age 55^+ and Gender Male, is significant for α =0.05, there is evidence that age and gender act as modifiers of effect on the relationship between ALS and contact practice. This result is in agreement with that obtained by the Breslow and Day test.

2) Through this model the odds of developing the disease for different age groups can be calculated comparing women or men practitioners and non-practitioners.

Age group $<$ 55 years:

Women Practicing contact sports

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} = e^{0.089} = 1.093$$

The odds of developing the disease in women $<$ 55 years, who practice contact sports, is 1.093 times higher, than for women in the same age group not practicing contact activity. This value agrees with the value obtained in stratum 1 tab.4.6.

Men Practicing contact sports

$$OR = \frac{e^{\beta_0+\beta_1+\beta_2+\beta_4}}{e^{\beta_0+\beta_2}} = e^{\beta_1+\beta_4} = e^{0.089+0.454} = 1.721$$

The odds of developing disease in men <55 years, who practice contact sports, is 1.721 times higher, than for men in the same age group not practicing contact activity. This value is in agreement with the value obtained in stratum 3 tab.4.6.

4. RESULTS

Age group 55^+ years:

Women Practicing contact sports

$$OR = \frac{e^{\beta_0+\beta_1+\beta_3+\beta_5}}{e^{\beta_0+\beta_3}} = e^{\beta_1+\beta_5} = e^{0.089-1.939} = 0.157$$

In woman 55^+ years, with intense contact activity the odds of getting the disease is 0.157 times lower than in woman of the same age without contact activity. The odds corresponds to that obtained for stratum 2 tab.4.6.

Men Practicing contact sports

$$OR = \frac{e^{\beta_0+\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_6+\beta_7}}{e^{\beta_0+\beta_2+\beta_3+\beta_6}} = e^{\beta_1+\beta_4+\beta_5+\beta_7} = e^{0.089+0.454-1.939+2.013} = 1.854$$

In men 55^+years, with intense contact activity the odds of getting the disease is 1.854 times higher than in men of the same age without contact activity. The odds corresponds to that obtained for stratum 4 tab.4.6.

For the previous model, age was considered categorized into two levels $<= 55$ and 55^+ in the remaining analyses, age will be considered in its continuous form. In this first approach, as the interest was the comparison of two methodologies, the variables to be inserted in the model, were defined from the beginning. Now, the problem will be approached, focusing on logistic regression, starting with the creation of univariate models, to select the variables for the multiple variable model.

Considering the empirical method, the first step consists in the creation of univariate models in order to select the candidate variables to enter the model, excluding those whose p-value is greater than **0.15**.

Table 4.8: Parameter estimates, Wald test-p-values, odd ratio (*OR*) and respective 95% confidence intervals of univariate models (Patients Vs Controls)

Variable		β	**p-value**	***OR***	**CI *OR* (95%)**
Age		0.03	$<< 0.001^{***}$	1.030	(1.024;1.037)
Gender					
	Female	Ref			
	Male	0.452	$<< 0.001^{***}$	1.572	(1.327; 1.862)
Smoking					
	Non-Smoking	Ref			
	Smoking	-0.195	0.119	0.822	(0.644; 1.052)
	Ex-Smoking	0.23	0.022^{**}	1.258	(1.035; 1.532)
TE		0.005	0.038^{**}	1.005	(1.001; 1.01)
Contact					
	No	Ref			
	Yes	0.478	0.002^{***}	1.613	(1.202; 2.184)

In addition, for the categorical variables, contingency tables of the observed values were constructed and analyzed for the dependent variable (Type = 0, 1) against the *k* levels of the independent variable in question, to assess whether for some of the variables, more than 20% of the cells have numbers less than 5. These tables are presented below.

Table 4.9: Contingency table of gender based on the type

Type	Gender	
	Female	Male
Patients	558	768
Controls	491	430

Table 4.10: Contingency table of smoking habits based on the type

Type	Smoking		
	Non Smoking	Smoking	Ex-Smoking
Patients	738	171	232
Controls	536	151	402

Table 4.11: Contingency table of contact sports practitioners based on the type

Type	Contact	
	Yes	No
Patients	152	1154
Controls	69	846

For this tables 4.9, 4.10, 4.11 there are no expected frequencies smaller than 5.

For the continuous variables, the linearity of logit was checked.

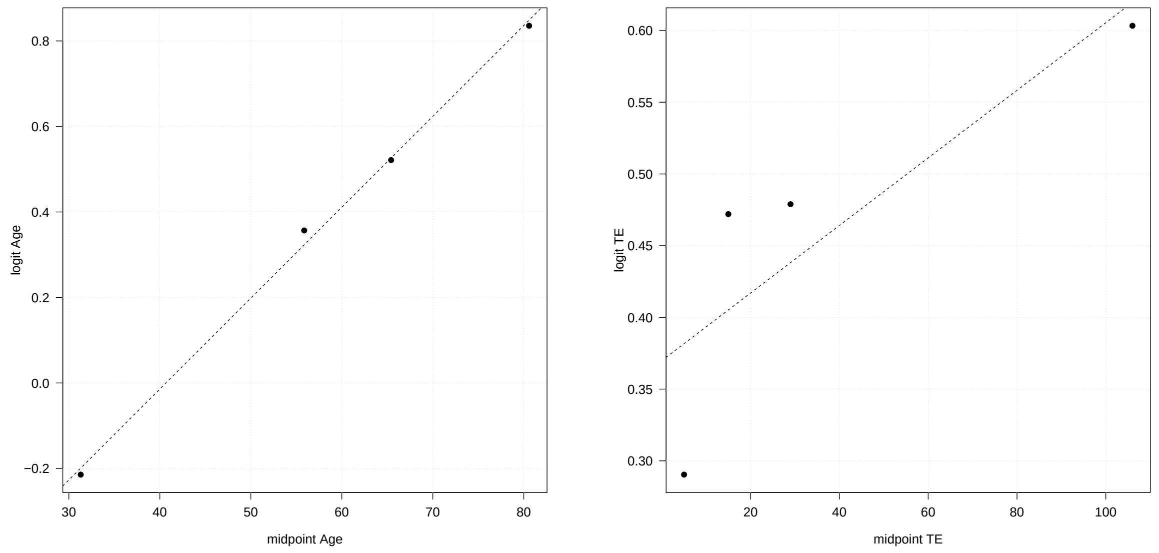


Figure 4.12: **a)** Plot estimated logit values for Age. **b)** Plot estimated logit values for TE.

Based on the results presented in 4.12 it is seen that age is linearly associated with ALS outcome in logit scale. The variable TE is not linear and might need some transformations, for example a spline function.

Table 4.12: AIC values and non-linearity tests for a model with different forms TE.

	TE	
	AIC	**p-value**
Linear	2720	
Restricted cubic splines		
knots		
3	2722	0.99
4	2723.7	0.859
5	2725.1	0.818
6	2725	0.547
7	2727.9	0.83

Based on the results presented in the table 4.12, for the variable `TE`, the linear form can be considered, as the linear model is the one with the smallest AIC. When looking at the p-values of the likelihood ratio test between the various non-linear models and the linear model, it is seen that these do not differ significantly.

As shown in table 4.8, all variables are candidates to enter in our model, however, in order to avoid problems of multicolinearity, in relation to smoking habits, it was decided to insert only the information regarding the load (TE).

The first model created, had the explanatory variables Contact, Age, Gender and TE, obtaining the first binary logistic regression model (4.2), as presented below:

Model II:

$$\ln\frac{p}{1-p} = 1.62 + 0.41 ContactYes_i + 0.373 GenderMale_i + 0.02 Age_i - 0.00009 TE_i \tag{4.2}$$

Table 4.13: Parameter estimates, Wald test-p-values, odd ratio (*OR*) and respective 95% confidence intervals of Model II without influent observations (Patients Vs Controls)

Variable		β	p-value	*OR*	CI *OR* (95%)
Intercept		1.62	$<<0.001^{***}$	0.198	(0.13;0.3)
Contact					
	No	Ref			
	Yes	0.41	0.017*	1.503	(1.08;2.114)
Gender					
	Female	Ref			
	Male	0.371	$<0.001^{***}$	1.45	(1.197;1.757)
Age		0.03	$<<0.001^{***}$	1.03	(1.023;1.037)
TE		9.269e-05	0.971	1	(0.995,1.005)

The next step was the insertion of interactions, although tobacco exposure was not statistically significant, we chose to keep it in the model. Initially, these interactions were tested Contact×Gender, Contact×Age and Contact×TE.

Model III

$$\begin{gathered}\ln\frac{p}{1-p} = -1.6 - 0.591 ContactYes_i + 0.323 GenderMale_i + \\ 0.03 Age_i - 0.002 TE_i + 1.431 ContactYes_i \times GenderMale_i + \\ -0.008 ContactYes_i \times Age_i + 0.021 ContactYes_i \times TE_i\end{gathered} \tag{4.3}$$

Table 4.14: Parameter estimates, Wald test-p-values for model III (Patients Vs Controls)

Variable		β	p-value	*OR*	CI *OR* (95%)
Intercept		-1.6	$<< 0.001^{***}$	0.202	(0.13, 0.307)
Contact					
	No	Ref			
	Yes	-0.591	0.515	0.554	(0.089, 3.173)
Gender					
	Female	Ref			
	Male	0.323	0.001^{**}	1.382	(1.136, 1.681)
Age		0.03	$<<0.001^{***}$	1.03	(1.023, 1.038)
TE		-0.002	0.534	0.998	(0.993, 1.004)
ContactYes× GenderMale		1.431	0.005^{**}	4.186	(1.586, 12.018)
ContactYes× Age		-0.008	0.545	0.992	(0.965, 1.019)
ContactYes×TE		0.021	0.042^{*}	1.021	(1.002, 1.044)

The `Contact × Age` interaction for α =5% is not statistically significant. Therefore this interaction was removed from the model. Comparing the models with (4.3) and without (4.4) interaction it does not differ significantly (p-value=0.514), so the following model was followed:

Model IV

$$\ln\frac{p}{1-p} = -1.576 - 1.056 ContactYes_i + 0.322 GenderMale_i + 0.029 Age_i$$
$$-0.002 TE_i + 1.435 ContactYes_i \times GenderMale_i + 0.02 ContactYes_i \times TE_i \tag{4.4}$$

Table 4.15: Parameter estimates, Wald test-p-values for model IV (Patients Vs Controls)

Variable		β	**p-value**	***OR***	**CI *OR* (95%)**
Intercept		-1.576	$<< 0.001^{***}$	0.207	(0.136, 0.313)
Contact					
	No	Ref			
	Yes	-1.056	0.03^{*}	0.348	(0.125, 0.872)
Gender					
	Female	Ref			
	Male	0.322	0.001^{**}	1.381	(1.135, 1.678)
Age		0.029	$<<0.001^{***}$	1.03	(1.023, 1.037)
TE		-0.002	0.545	0.998	(0.993, 1.004)
ContactYes× GenderMale		1.435	0.005^{**}	4.201	(1.585, 12.107)
ContactYes×TE		0.02	0.051	1.02	(1.002, 1.043)

By the results obtained, the amplitude of the CIs acts as a warning signal. For the Contact × GenderMale interaction, the odds ratio CI of [1.585, 12.107], the amplitude of these range is high.

The relationship between the practice of contact sports and the disease is different between genders. In women, the practice of contact sports appears as protective against the disease, while in men it appears as a risk factor. This difference between genders seems to be influenced by the small number of women who practice contact sports. The number of female contact practitioners is low. Only 25 women fits in this class and 7 are sick. Considering the average age of the female patients in this study - 63.5 years, this can be an expected difference, as in this generation physical activity with the intensity level and the type of sports included in this work, was not common for the female population. Therefore, it was decided to only consider men.

The first model created, had as explanatory variables: Contact, Age and TE. Through these, the following results were obtained:

Model V

$$\ln \frac{p}{1-p} = -0.548 + 0.616 ContactYes_i + 0.017 Age_i + 0.002 TE_i \tag{4.5}$$

Table 4.16: Parameter estimates, Wald test-p-values for model V (Patients Vs Controls)

Variable		β	p-value	*OR*	CI *OR* (95%)
Intercept		-0.548	0.062	0.578	(0.324, 1.028)
Contact					
	No	Ref			
	Yes	0.616	0.001**	1.851	(1.286, 2.707)
Age		0.017	0.0006***	1.017	(1.007, 1.027)
TE		0.002	0.599	1.002	(0.996, 1.008)

The next step was the insertion of interactions, although tobacco exposure was not statistically significant, it was kept it in the model. Initially, interactions were tested Contact×Age and Contact×TE.

Model VI

$$\ln\frac{p}{1-p} = -0.44 - 0.22ContactYes_i + 0.015Age_i$$
$$-0.001TE_i + 0.009ContactYes_i \times Age_i + 0.024ContactYes_i \times TE_i \tag{4.6}$$

Table 4.17: Parameter estimates, Wald test-p-values for model VI (Patients Vs Controls)

Variable		β	p-value	*OR*	CI *OR* (95%)
Intercept		-0.44	0.159	0.644	(0.348, 1.187)
Contact					
	No	Ref			
	Yes	-0.22	0.799	0.802	(0.145, 4.373)
Age		0.015	0.003**	1.015	(1.005, 1.026)
TE		-0.001	0.733	0.999	(0.992, 1.006)
ContactYes× Age		0.009	0.534	1.009	(0.98, 1.04)
ContactYes×TE		0.024	0.045*	1.024	(1.003, 1.051)

The Contact × Age interaction for α =5% is not statistically significant. Therefore this interaction was removed from the model. Comparing the models with and without interaction it does not differ significantly (p-value=0.566), so the following model is:

Model VII

$$\ln\left(\frac{p}{1-p}\right) = -0.507 + 0.299 ContactYes_i + 0.017 Age_i -$$

$$0.001 TE_i + 0.025 ContactYes_i \times TE_i \tag{4.7}$$

Table 4.18: Parameter estimates, Wald test-p-values for model VII (Patients Vs Controls)

Variable		β	**p-value**	***OR***	**CI *OR* (95%)**
`Intercept`		-0.507	0.084	0.602	(0.338, 1.07)
`Contact`					
	`No`	Ref			
	`Yes`	0.299	0.189	1.349	(0.867, 2.124)
`Age`		0.017	$<0.001^{***}$	1.017	(1.007, 1.027)
`TE`		-0.001	0.7	0.999	(0.992, 1.005)
`ContactYes×TE`		0.025	0.032^{*}	1.026	(1.004, 1.052)

Considering the defined objective - to evaluate the relationship between the practice of contact sports and ALS, it was found that the odd of developing the disease is 1.3 times higher in practitioners, compared to non-practitioners for non-smokers (TE=0). Regarding potential confounders, the interaction `Contact×TE` is statistically significant.

The difference between the odds ratio for practitioners vs non-practitioners is 1.03 in two homogeneous loads that differ in 1 pack-year. That means an increase of approximately 3% in the odds of developing the disease in practitioners when compared to non-practitioners. For each year more in age, the chance of developing the disease is 1.02 times higher.

Next, diagnosis of the model is presented.

Table 4.19: Studentised Pearson residuals values(rpsi), leverage (h$_{ii}$) and Cooks distance (Δ_i)

ID	rps$_i$	h$_{ii}$	$\Delta\beta_i$
726	-2.03	0.017	0.019
767	-2.06	0.0153	0.018
957	-1.225	0.051	0.01
960	-1.065	0.1	0.014
1001	-2.09	0.021	0.027

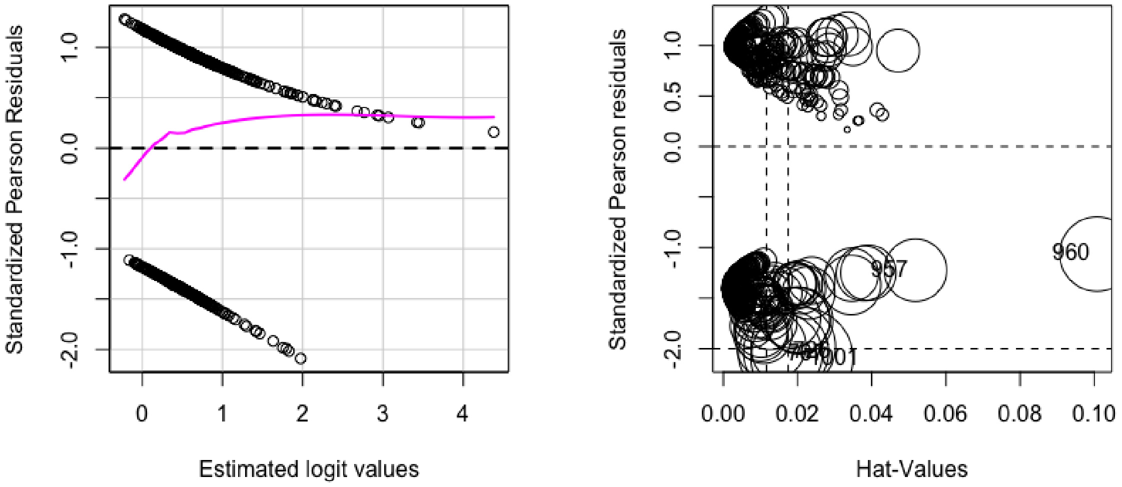


Figure 4.13: **a)**Plot of standardized Pearson residuals as a function of estimated logit values. **b)** Plot of hat-values, Studentized residuals, and Cook's distances for regression.

To evaluate the fit quality of the model, the Stukel score test was applied, obtaining a p-value of 0.671. So, the model can be considered well adjusted to the data.

The residues were then analyzed to verify if the good fit was supported in the entire data set. Through the analysis of the left graph of fig.4.13, it was found that two standard Pearson residues had an absolute value greater than 2, so outliers or influential observations may be possible. It was also observed that the downward curve incorporated in the graph resulted approximately in a horizontal line with the intersection, except for the most extreme logit values, suggesting that the model is correct and that there are no significant outliers.

Observations 726, 767, 957, 960 and 1001 were identified as possible influential observations. When the 5 observations were removed individually, there were no relevant changes in the model coefficients. So, they could be kept in the model.

Validation of Predicted Values

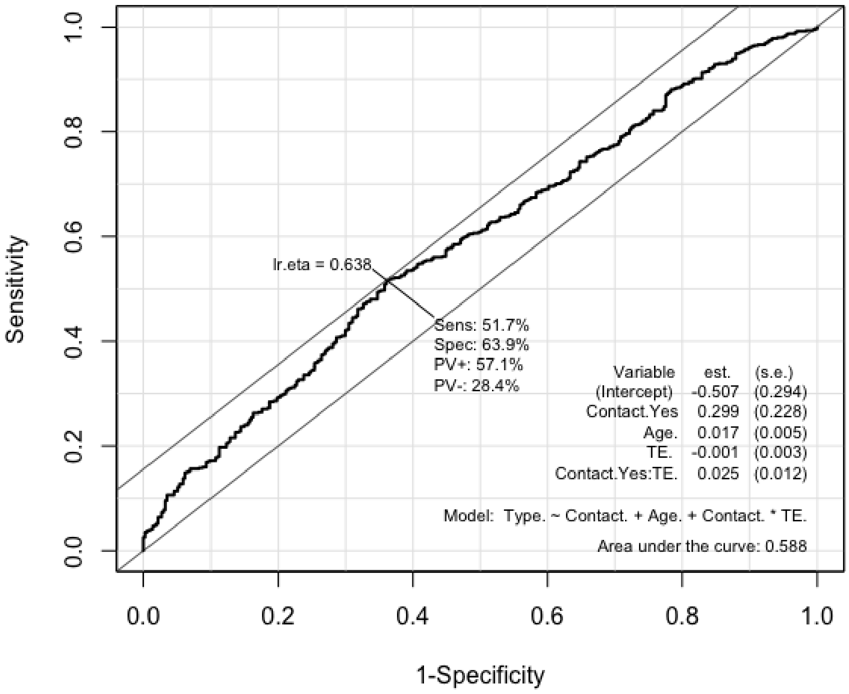


Figure 4.14: ROC Curve for model VII

From the analysis of the ROC curve, it seems that the model has a relatively low predictive capacity and it has an area under the curve of 0.588. Ideally a value close to 1 shoud be obtained, which is rare when working with this type of data.

4.3 Evaluate association between the age of early diagnosis and the practice of high-intensity football

After working on the first objective - to assess the relationship between the practice of contact modalities and ALS and the existence of variables that may interfere in this relationship, the nest step was to assess the relationship between the age of diagnosis and the practice of high intensity football - the second objective. One-way ANOVA was used.

Table 4.20: Means, standard deviance and frequencies for diagnostic age at different football levels for the Portuguese sample.

	Diag.Age		
Football levels	Mean	Std.Deviacion	Frequency
No	64.53	12.9	385
Mild	65.41	15	5
Intense	56.57	11.4	36

Table 4.21: Means, standard deviance and frequencies for diagnostic age at different football levels for general sample.

	Diag.Age		
Football levels	Mean	Std.Deviacion	Frequency
No	61.09	13.4	1159
Mild	61.76	13.3	10
Intense	58.74	11.9	134

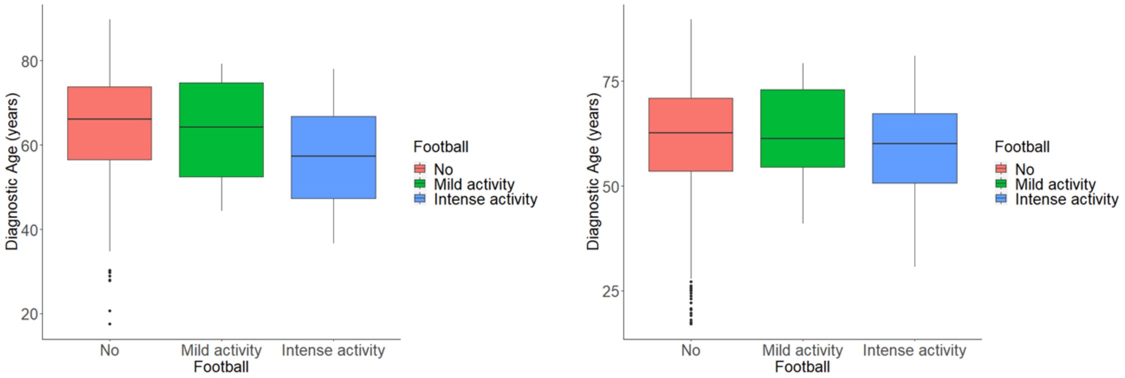


Figure 4.15: Parallel boxplots for **a)** Diagnostic Age for Portugal population and **b)** Diagnostic Age for general population by level of football

.

From the boxplots in fig.4.15, some preliminary observations can be made. For both cases, Portuguese fig.4.15a) and general fig.4.15b) it can be seen that in the ”Don’t play football” level there are some outliers but in the other levels not. For the Portuguese sample, visually there seems to be differences between the 3 intensity levels. Aiming to see if these differences are statistically significant or not, a one-way ANOVA analyse was made.

Portuguese Population

Table 4.22: The one-way ANOVA results

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Football	2	2097	1048.7	6.423	0.00179***
Residuals	423	69067	163.3		

Based on the information in table 4.22, the hypothesis that the α_i level increments are all null is rejected. So, the hypothesis that the mean age of diagnosis is equal at all intensity levels had to be rejected.

In conclusion, there is a difference in the mean age of diagnosis between at least one of the intensity levels. Still, it is necessary to identify between what levels these differences exist.

Before identifying for which levels there are differences in diagnostic age, the assumptions of the model were evaluated. As the groups are unbalanced, particular attention to the assumption of homogenity of variance, must be given. For the validation of the assumptions, the plot representation of the residuals was chosen.

4.3 Evaluate association between the age of early diagnosis and the practice of high-intensity football

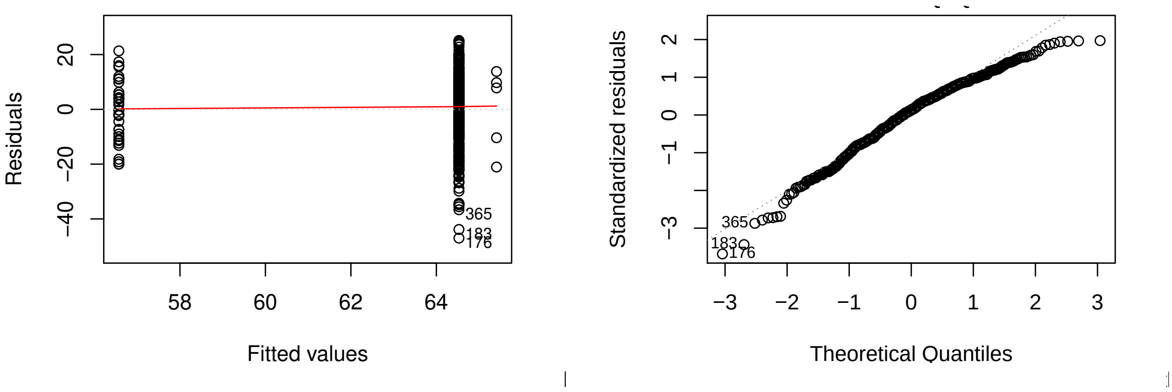


Figure 4.16: **a)** Residuals vs fitted plot **b)** QQ plot of residuals.

From the analysis of fig.4.16, no apparent pattern was visible, so it looks like the constant variance assumption was satisfied. Regarding normality, with the exception of the lower and upper extremities, it seems that the assumption of normality was also verified. Thus we could rely on the results obtained and move on to the multiple comparisons test to see which of the levels differ.

Multiple Comparisons

Table 4.23: 95% Tukey-Kramer confidence intervals

	Difff	Lower	Upper	p adj
No-Mild activity	-0.88	-15.713	13.954	0.987
No-Intense activity	7.959	2.215	13.702	0.001
Mild-Intense	8.838	-6.88	24.566	0.317

The multiple comparisons test, demonstrate (tab.4.23) that there are only differences in the average age of diagnosis between non-football players and intensive football players, those that practice intense football have an age at diagnosis that is 8 years younger than those who do not practice.

General Population

Table 4.24: The one-way ANOVA results

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Football	2	667	333.3	1.892	0.151
Residuals	1300	229055	176.2		

From the result of the ANOVA test, a high p-value was obtained. So, when considering the entire population, there was no evidence of differences in the average ages of diagnosis between the different intensity levels, as it was when considering exclusively the data regarding the Portuguese population.

Assumption Evaluation

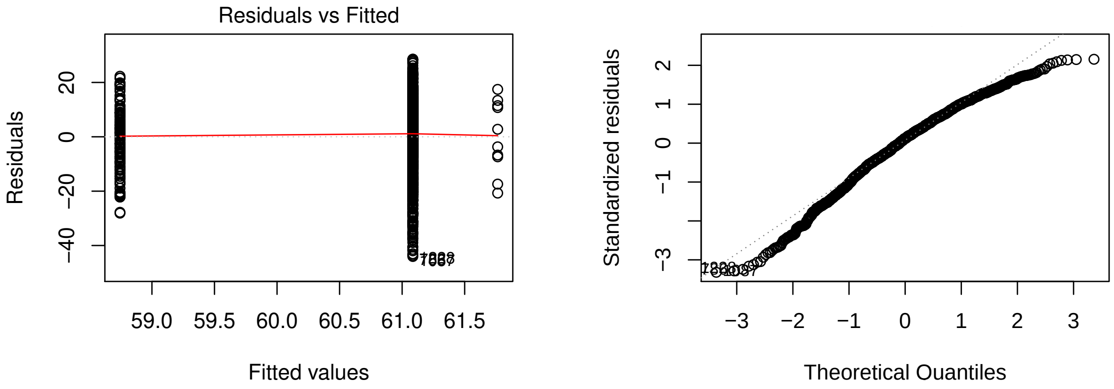


Figure 4.17: **a)** Residuals vs fitted plot **b)** QQ plot of residuals.

Residuals considering the total population behave similarly as when considering only the Portuguese population. By analyzing the fig 4.17, regarding the equality of variances, as there is not a pattern, the homogeneity of variances can be assumed. In relation to normality, especially at the extremities, the points deviate from the standard line, however, it is not significant to reject the results obtained considering the robustness of ANOVA in relation to deviations from normality.

4.4 Secondary goals

To assess the association between the delay in diagnosis and the age of the first symptoms and the type of onset, the option was to create a linear model with the `Diagnostic Delay` as the dependent variable and as independent variables the type of onset (`Onset`) and the age of the first symptoms (`Age.1st.Sym`).

Getting the following model:

$$y = 35.5 - 6.794 OnsetBulbar - 0.273 Age.1st.Sym \tag{4.8}$$

Before moving to the interpretation of the model, the assumptions of normality and equality of variance (homoscedasticity) were evaluated, since this is a necessary premise for all the inferential process that will be performed in the model.

To evaluate the assumption of equality of variance (homoscedasticity) and normality of errors in linear regression, the option was to plot the residuals.

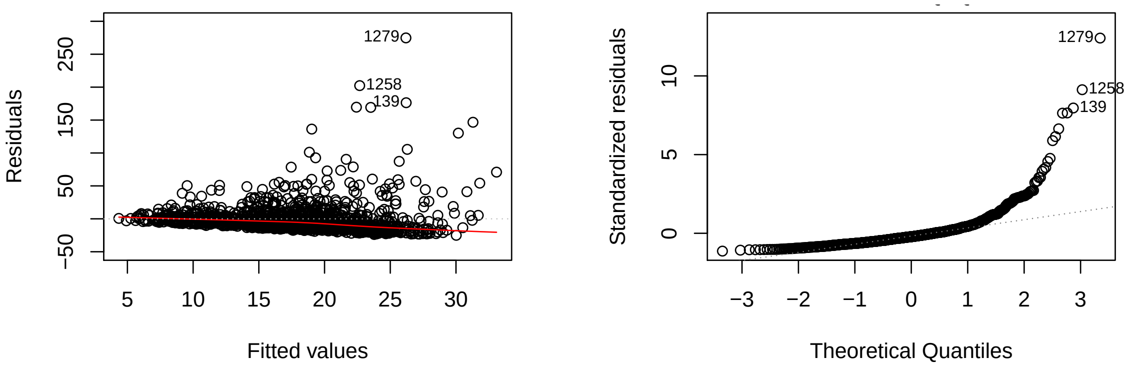


Figure 4.18: **a)** Residuals vs fitted plot **b)** QQ plot of residuals.

From the analysis of the plots in figure 4.18, it is visible that the assumptions of homoscedasticity and the normal distribution of errors are not met. Therefore, the values obtained in this model could not be considered.

One of the possible and most used solutions in these cases is the logarithmic transformation. The chosen solution was to transform the dependent variable diagnostic delay, reaching the following model:

$$\ln(Diagnostic.Delay) = 2.893 - 0.336 OnsetBulbar - 0.007 Age.1st.Sym \tag{4.9}$$

As done in the untransformed model, before proceeding to the interpretation, the fulfillment of the assumptions was checked.

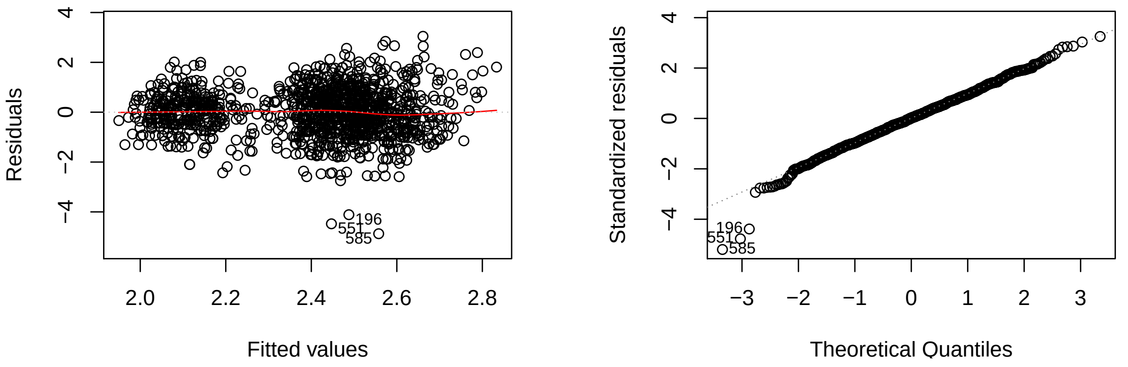


Figure 4.19: **a)** Residuals vs fitted plot **b)** QQ plot of residuals.

No pattern is visible on the residuals graph fig 4.19 a), which suggests compliance with the assumption of homoscedasticity. In fig 4.19 it is also seen that b) all points fall approximately at the reference line except the most extreme values. Therefore, it is assumed that errors follow a normal distribution.

Observations with Id 196, 551 and 585 corresponds to outliers candidates $|Ti| > 3$.

After checking the assumptions, the existence of influential observations was evaluated. The observations with Id 39, 551, 585, 650, 785 and 1105 were pointed to as influential observations, so they were removed one by one to check the effects on the model coefficients.

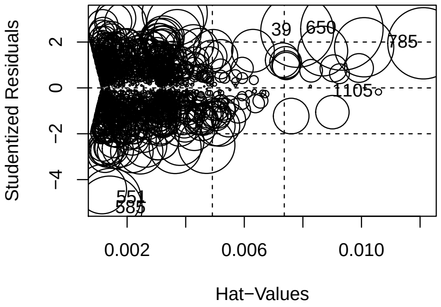
By removing the observations one by one, no changes greater than 10% in the coefficients of the model were observed and the same happens when removed in bulk, so they were kept in the model.

Figure 4.20: Representation of influential observations

(a) Studentised Pearson residuals values(rpsi), leverage (h_{ii}) and Cooks distance (Δ_i)

ID	rps$_i$	h$_{ii}$	$\Delta\beta_i$
39	2.485	0.008	0.016
551	-4.819	0.001	0.011
585	-5.257	0.001	0.013
650	2.569	0.009	0.021
785	1.945	0.012	0.015
1105	-0.179	0.011	0.0001

(b) Plot of hat-values, Studentized residuals, and Cook's distances for model.



The coefficients of the model are presented in table 4.25

Table 4.25: Parameter estimates, standard errors, p-values and % Confidence Interval

Variable	β	Std. Err.	**p-value**	**CI (95%)**
`Intercept`	2.893	0.121	$<< 0.001^{***}$	(2.657,3.13)
`Onset Bulbar`	-0.336	0.062	$<< 0.001^{***}$	(-0.458,-0.213)
`Age.1s.Sym`	-0.007	0.002	0.001^{***}	(-0.011,-0.003)

Being a log-linear model the interpretation was made by taking into account the median and not the average, due to the characteristics of this type of models.

With the intercept (the expected mean for ln(diagnostic.delay)) for limb onset when age.1st.symptom is equal to zero), the diagnostic delay was approximately 30% ($\exp(-0.336) = 0.715$) lower for the patients with bulbar onset than for the patients with limb onset. For the age.1st.symptom variable, by the increase of one year in the age of the onset of the first symptom, a decrease of about 0.7% in the diagnostic delay ($\exp(-0.007) = 0.993$), was expected.

This model is useful for determining the relationship between variables, however it should not be considered for the purpose of prediction, since it has a very low coefficient of determination (0.041).

To evaluate the characteristics associated with different values of ALS Fuction Rating Rate of decay (ALSFRS.R), the same methodology was applied. Taking as response variable ALSFRS.R and as predictors the `Diagnostic delay`, `Onset`, `Contact`, `Gender`, `Age` and `TE`.

The predictors of this model were chosen considering two criteria: first the group of variables that may be related to the different levels of progression, namely `Diagnostic delay` and `Onset`, then a second group of variables `Contact`, `Gender`, `Age` and `TE`, were inserted into the model to evaluate if after the disease diagnosis, the considered risk factors continue to have influence in the development of the disease and if they are related to the different levels of the disease progression.

The following model was obtained:

$$\begin{aligned} ALSFRS.R = 0.516 + 0.009Age - 0.006GenderMale + 0.001TE - \\ 0.22ContactYes - 0.01Diagnostic.Delay + 0.09OnsetBulbar \end{aligned} \tag{4.10}$$

After creating the model, once again the assumptions were evaluated based on the graphical representation of the residuals. By the information in fig.4.21, it can be seen that the assumption of equality of variances and the normal distribution of errors is not fulfilled. By the plot on the right, it is seen that there are a set of points that deviates from the trend line.

As in the previous case, logarithmic transformation can be a good alternative. However, in this case there were 8 null observations for y, so, the logarithmic transformation could not be applied directly. These observations concerns to individuals who scored the maximum value when asked about their functional level.

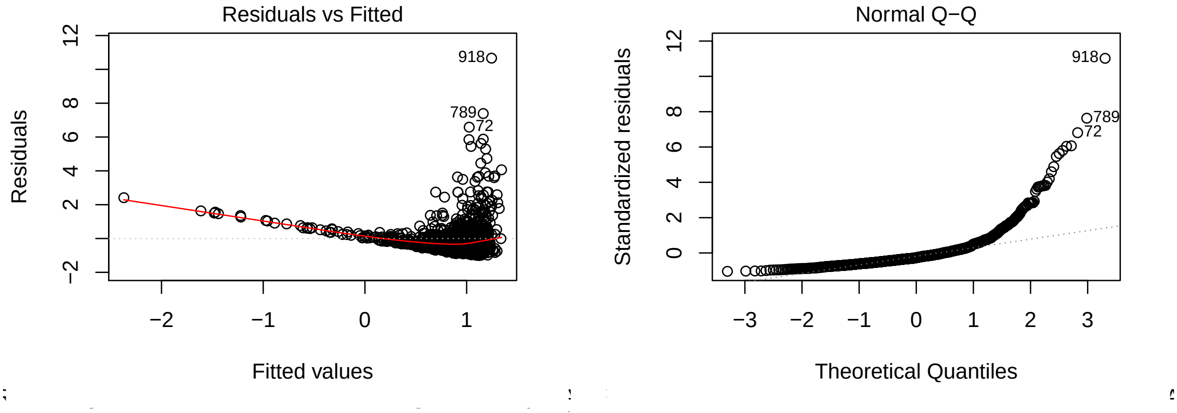


Figure 4.21: **a)** Residuals vs fitted plot **b)** QQ plot of residuals.

Various techniques for log transformation with zeros are described. In this case, as there are only 8 observations, out of a total of 1326, the option was to withdraw these observations, as their removal would not have much influence on the estimation of the model parameters.

By applying the logarithmic transformation to the response variable, the following model was obtained:

$$\begin{aligned} \ln(ALSFRS.R) = -0.74 + 0.008Age - 0.151GenderMale + 0.002TE \\ -0.12ContactYes - 0.023Diagnostic.Delay + 0.136OnsetBulbar \end{aligned} \tag{4.11}$$

After applying the logarithmic transformation, the assumptions of the model were checked.

Based on fig.4.22 it turns out that even after the transformation of the predictor variable, there were still problems. The way the residuals are exposed suggests a curvature in the relationship between the variables, and there are also some residues larger than $|Ti| > 3$. Regarding normality, there was only a small deviation for the most extreme values.

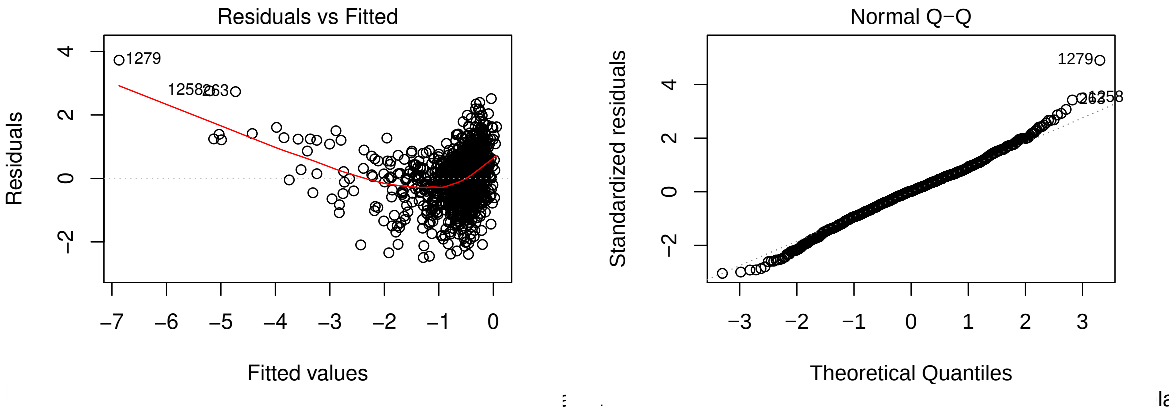


Figure 4.22: **a)** Residuals vs fitted plot **b)** QQ plot of residuals.

The behavior of the continuous variables present in the `Diagnostic Delay` and `Age` model was verified in fig.4.23. For the variable `Age`, no pattern is verified, for the `Diagnostic Delay` there is a curvature, which leds to consider the application of the logarithmic transformation in this variable.

The transformation was tested considering the univariate model and there was an improvement in the distribution of residiuals. The transformed variable was inserted in the model.

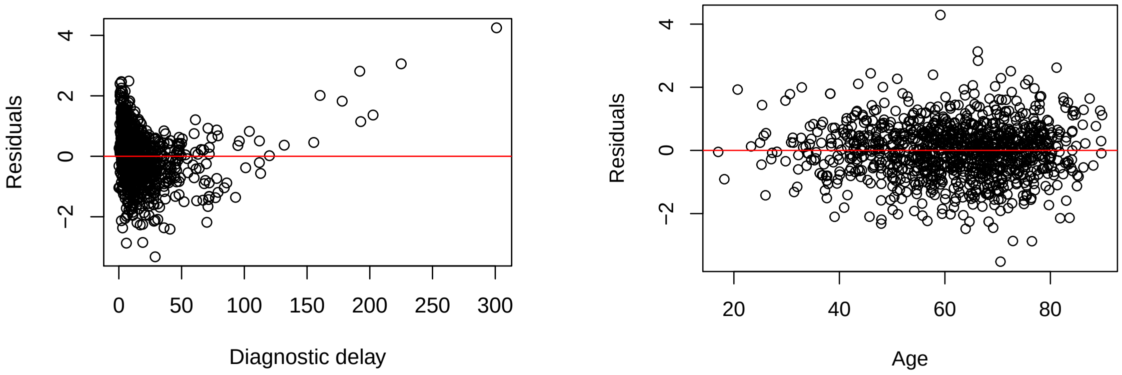


Figure 4.23: Residual plot

By applying the logarithmic transformation to the `Diagnostic Delay` variable, the following model was obtained:

$$\begin{aligned} \ln\left(ALSFRS.R\right) &= 0.453 + 0.01Age - 0.218GenderMale + 0.002TE \\ &-0.08ContactYes - 0.69\ln\left(Diagnostic.Delay\right) + 0.045OnsetBulbar \end{aligned} \tag{4.12}$$

After applying the logarithmic transformation, the assumptions of the model was checked.

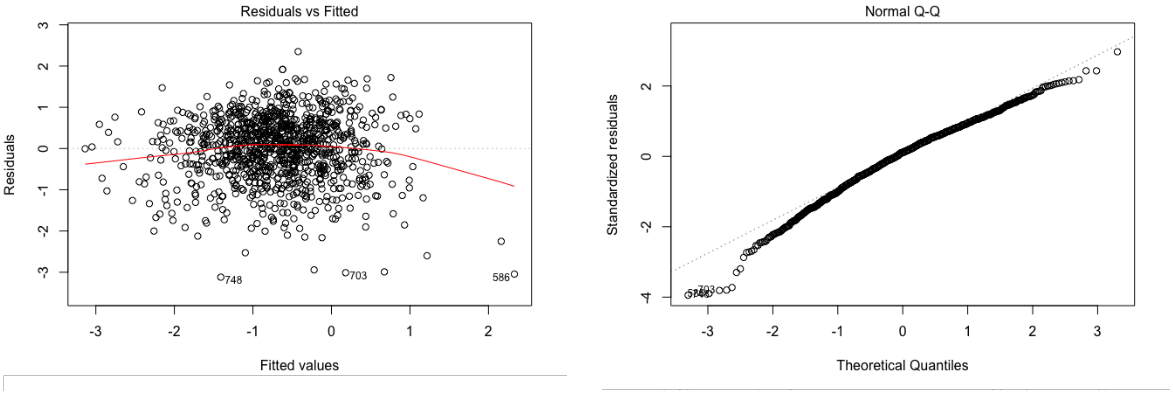


Figure 4.24: **a)** Residuals vs fitted plot **b)** QQ plot of residuals.

No pattern is visible on the residuals graph fig 4.24 a), which suggests compliance with the assumption of homoscedasticity. In fig 4.24 it is also seen that b) most points fall approximately on the reference line, except for the most extreme values. Therefore, it is assumed that errors follow a normal distribution. Particular attention should be paid to observations with Id 748, 703, 586 .

After checking the assumptions, the existence of influential observations was evaluated. The observations with Id 227, 586, 703, 840, 1039 and 1279 were pointed to as influential observations, so they were removed one by one to check the effects on the model coefficients.

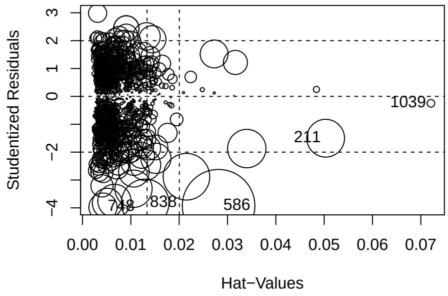
By removing the observations one by one, no changes greater than 10% in the coefficients of the model were observed and the same happens when removed in bulk, so they were kept in the model.

Figure 4.25: Representation of influential observations

(a) Studentised Pearson residuals values(rpsi), leverage (h_{ii}) and Cooks distance (Δ_i)

ID	rps_i	h_{ii}	$\Delta\beta_i$
211	-1.59	0.05	0.017
586	-3.925	0.028	0.063
748	-3.973	0.004	0.009
838	-3.82	0.013	0.027
1039	-0.25	0.072	0.0001

(b) Plot of hat-values, Studentized residuals, and Cook's distances for model.



Then, the model was interpreted. It is important to note that due to the characteristics of this type of models, the interpretation of the model parameters is done for the median and not for the average.

Table 4.26: Parameter estimates, standard errors, p-values and % Confidence Interval

Variable		β	Std. Err.	**p-value**	**CI (95%)**
Intercept		0.453	0.14	0.011 **	(0.179, 0.727)
Age		0.01	0.002	<<<0.001***	(0.006, 0.013)
Gender	Female	Ref			
	Male	-0.218	0.05	0.001***	(-0.322, -0.114)
TE		0.002	0.001	0.065	(-0.001, 0.004)
Contact	No	Ref			
	Yes	- 0.079	0.08	0.322	(-0.236, 0.077)
ln(Diagnostic.Delay)		-0.69	0.0261	<<<0.001***	(-0.742, -0.64)
Onset	Limbs	Ref			
	Bulbar	0.045	0.057	0.437	(-0.112, 0.102)

The variables Contact, TE and Onset are not significant at the level $\alpha = 5\%$ however, they were maintained in the model, because they are biologically relevant and constitute the main focus of the study. Based on the results obtained, for each additional year age, there is an increase of approximately ($\exp(0.001) = 1.001$) 0.1% for the progression ratio, comparing men and women the values of ALSFR.R are approximately ($\exp(-0.21) = 0.8$) 20% lower in men compared to women. An increase of one percent in the diagnostic delay, corresponds to an decrease of approximately ($1.01^{-0.69} = 0.993$) 0.7% of progression levels.

This model is useful for determining the relationship between variables, however it should not be considered for the purpose of prediction, since it has a low coefficient of determination (0.45).

Finally, to assess whether there are differences between the different levels of disease progression, multinomial logistic regression models were used.

A multinomial logistic regression with the variable Progression Group with the 3 categories already indicated (slow, neutral and fast), was also performed.

The slow category was defined as the comparison class. Thus, two groups of results were obtained. One resulting from the comparison between slow progression vs neutral and the other from the slow progression vs fast.

As for logistic regression, in these cases the first step is to create univariate models for selecting the variables to include in the model.

Based on the univariate analysis, the results presented in table 4.27 were obtained by comparing slow vs neutral and slow vs fast progression groups.

Table 4.27: Parameter estimates, odd ratio (*OR*) and respective 95% confidence intervals of univariate multinomial models (Neutral Vs Slow and Fast vs Slow), F represent results for comparison Fast vs Slow and N represent results for comparison Neutral vs Slow

Variable		β	**p-value**	***OR***	**CI *OR* (95%)**
`Age`					
	`Age [N]`	0.013	$<<< 0.001^{***}$	1.014	(1.004; 1.023)
	`Age [F]`	0.031	$<<< 0.001^{***}$	1.032	(1.017; 1.046)
`Gender`					
	`Female`	Ref			
	`Male [N]`	-0.311	0.011^{**}	0.732	(0.576;0.932)
	`Male [F]`	-0.632	$<0.01^{***}$	0.532	(0.379;0.746)
`TE`					
	`TE [N]`	0.003	0.317	1.003	(0.997,1.01)
	`TE [F]`	0.006	0.169	1.006	(0.998,1.014)
`Onset`					
	`Limbs`	Ref			
	`Bulbar [N]`	0.807	$<<0.01^{***}$	2.24	(1.689;2.974)
	`Bulbar [F]`	0.95	$<<0.01^{***}$	2.586	(1.768;3.781)
`Diagnostic delay`					
	`Diagnostic delay [N]`	-0.096	$<0.01^{***}$	0.908	(0.997,1.01)
	`Diagnostic delay [F]`	-0.329	$<0.01^{***}$	0.719	(0.997,1.01)
`Contact`					
	`No`	Ref			
	`Yes [N]`	0.036	0.842	1.04	(0.726;1.48)
	`Yes [F]`	-0.86	0.013^{**}	0.423	(0.214;0.837)

Based on univariate models, all variables are candidates for entering the model, with the exception of tobacco exposure `TE`. So, the model was recalculated without this variable and the results presented in Table 4.28.

Table 4.28: Parameter estimates, odd ratio (*OR*) and respective 95% confidence intervals of multinomial model (Neutral Vs Slow and Fast Vs Slow), F represent results for comparison Fast vs Slow and N represent results for comparison Neutral vs Slow

Variable		β	**p-value**	***OR***	**CI *OR* (95%)**
`Age`					
	`Age [N]`	0.018	$<< 0.001^{***}$	1.018	(1.007;1.03)
	`Age [F]`	0.046	$<< 0.001^{***}$	1.05	(1.029;1.066)
`Gender`					
	`Female`	Ref			
	`Male [N]`	-0.323	0.035^{**}	0.724	(0.536;0.977)
	`Male [F]`	-0.712	0.002^{***}	0.49	(0.309; 0.778)

Onset					
	Limbs	Ref			
	Bulbar [N]	0.292	0.756	1.339	(0.967;1.855)
	Bulbar [F]	-0.076	0.079	0.926	(0.571; 1.502)
Diagnostic delay					
	Diagnostic delay [N]	-0.099	$<<0.001^{***}$	0.905	(0.889,0.921)
	Diagnostic delay [F]	-0.34	$<<0.001^{***}$	0.712	(0.672; 0.754)
Contact					
	No	Ref			
	Yes [N]	0.186	0.4	1.205	(0.78;1.86)
	Yes [F]	-0.675	0.122	0.509	(0.217; 1.196)

In this simpler model, the variables `Onset` and `Contact` loose their statistical significance. Still, they are important for the purposes of the study and no further simplifications were attempted.

Based on the results shown, it can be said that for each year more in age, the odds of being part of the neutral progression group is 1.05 times higher and the odds of being part of the fast progression group is 1.02 times higher. For the gender, the odds of being part of the neutral progression group is 0.51 times lower in men compared to women and for the fast progression group the odd of becoming part of it, is 0.28 times lower in men compared to women.

For each month more in the delay of diagnosis, the odds of being part of the neutral progression group is 0.095 times lower. For the group of fast progression for each month more in the delay of diagnosis the chance to be part of this group is 0.288 times lower.

Chapter 5

Discussion and Conclusion

Recently, the relationship between physical activity and sports-related trauma is being discussed. In this study, an association between the practice of high intensity contact modalities and ALS, was found. The results are aligned with the ones published by Blecher et al. (2019), which refers that intense activity and sports associated with successive head and neck trauma increases the risk of developing ALS. This is a systematic review of 16 previous studies. The theme fits the main objective of this thesis - contact sports as a risk factor for amyotrophic lateral sclerosis. The outcome was the incidence of ALS or mortality associated with ALS and exposure to any organised competitive sport, professional or non-professional, defined a priori as the sports with the greatest exposure of players to repetitive trauma to the head and back or cervical spine: American football, soccer, hockey, boxing, rugby. Compared to the 16 reviewed studies, ours is one of the largest in terms of the sample size.

Significance for the practical interactions between contact sport, gender and age when categorized, was also found. However, as already mentioned in the results, this interaction may be associated with the reduced number of females practitioners of contact modalities, with a high intensity level. That is why, for the final model, the male population was exclusively considered.

Regarding smoking habits, contrary to what is described in other studies, where smoking and the number of cigarettes per day is identified as a risk factor for ALS (Wang et al., 2012), in this study, although there is an association between smoking and the onset of the disease, this relationship is quite small. However significance for the Contact × TE interaction was found, an interaction that until now has not been identified in previous published results.

Another goal proposed was to check the relationship between football practice and the age of diagnosis. We believe that an early age of diagnosis may be associated with an high intensity football practice.

Considering only the Portuguese population, differences in the age of diagnosis between high intensity practitioners and low intensity or non-practitioners were found. However, this difference was not verified when considering the whole population.

The difference between the Portuguese population and the whole population may be related with the fact that the Portuguese sample is more homogeneous, because it was curated by the same ALS center.

Other studies points football as a risk factor for the disease and several known cases of former players with this problem, but still, the relationship between early diagnosis and intense football practice has not been explored yet.

From the results obtained, the limbs onset and an early diagnosis age are associated with an increase in diagnostic delay. The results presented in the work of Paganoni et al. (2014) are in agreement with the ones found now, referring that limbs onset and advanced age are associated with increased diagnostic delay. In this case, advanced age gets a decrease of 0.7% in the diagnostic delay for each year more in age 1st.symp. However, in the same article the fact that older people resort more to doctors compared to younger ones, is highlighted. It can therefore be considered that for the age at which the peak of the disease is described, people turn to doctors often. So, this will increase the odd of getting a diagnosis earlier than younger people who go to doctors less often.

On the other hand, at an older age, as for example in some cases included in this database, closer to 80 years, the symptoms of the disease can be confused with other age-related problems, which contributes to an increase in the diagnostic delay.

In Kimura et al. (2006) a comparison between two slow-progressing and faster-progressing groups was made. It was concluded that in the fastest-progressing group, the delay to diagnosis was significantly shorter. When considering the ALSFRS-R model as the predictor variable, the results obtained in this work are in agreement with those presented by the author. Both concluded that an increase of one percent in the diagnostic delay, corresponds to a decrease of approximately 0.6 % in the ALSFRS-R.

When considering the multinomial model, comparing neutral versus slow, there's a 9.5% decrease in the odds of being part of the neutral progression group compared to the slow progression group for each extra month in the delay in diagnosis. When considering the fast vs slow comparison, the decrease is even steeper.

In the present study, significant differences were also found between `Sex` and `Age` in both neutral vs slow and fast vs slow comparison.

Methodological Issues

When using questionnaires to collect data, associated limitations like some lack of honesty in the answers, some unanswered questions, lack of coherence between answers and so on, are always expected. These limitations are difficult to control and acts as a source of error.

The reduced number of females' practitioners of contact modalities, with a high intensity level, is another limitation that may have compromised some of the obtained results.

For a large number of variables, the stratified analysis in case-control studies, is not the best one. The binary logistic regression model is a good alternative, for these cases. In this study, as there was not a large number of variables to test, both methodologies were used and the results were compared. As expected, results were equivalent.

With the exception of the model that served to compare the results obtained by the stratified analysis and the logistic regression models in which age was categorized into two groups, for the remaining analyses it was considered in a continuous way. One of the most common approaches, when working with continuous predictors, such as age, is to categorize it, an approach that can lead to loss of information (Gauthier, Wu, & Gooley, 2019). For example, considering the age-related dicthomization in the present case, in which age was classified as 55 or less years and more than 55, this dicthomization assumes that a participant aged 55, has the same association with the result, as for example, an 80 year old person, but different from a 30 year old participant, which for ALS we know is not quite, although the disease can strike at any age, symptoms most commonly develop between the ages of 55 and 75 (National Institute of Neurological Disorders and Stroke NIH, n.d.).

Citing the emblematic statement of Box (1979) "all models are wrong, but some are useful", it is important to note that the results presented here cannot be seen as a universal law, but as a support to better understand the existence of some associations, like the existence of a relationship between the practice of contact sports and ALS.

Conclusion

In conclusion, the main objective of this study was achieved - study the relationship between ALS and the practice of contact sports.

Considering the Portuguese population, differences in the age of diagnosis for high intensity football players was also found. When considering the entire population, differences were not found.

Regarding the delayed diagnosis, it can be said that the limbs onset and an early age of diagnosis are associated with an increase in the delayed diagnosis. Finally, considering the different progression groups, it is visible that slower progression groups correspond to a longer period until diagnosis.

The evaluation of the relationship between ALS and sport, namely the practice of contact sports, as assessed in this report, is currently being analysed in several studies, such as the work published by Blecher et al. (2019).

The significant interaction obtained between gender, age and contact modalities, seems also interesting, although results may be influenced by the small number of women practicing contact sports in this sample, as discussed before. As is expected that in a few years the difference between men and women in the practice of contact modalities will not be so pronounced, the development of a similar study in the future will be important. To confirm the existence of a relationship between ALS and the practice of high-intensity contact modalities, more research is also needed. This will be essential to bring athletes and sports federations attention to this problem.

Although the results presented in this study are not innovative, with the exception of the significance found for interactions between the practice of contact sports and tobacco exposure, they are aligned with previous studies, are updated and meet the need for more research in this field.

As far as we know, this is one of the largest international studies in the ALS field. As the results now achieved, can be a relevant input to reinforce and validate previous similar studies, they should also be published to support future research and to be a positive contribution for a better understanding of the disease.

Bibliography

Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, *54*, 979–985. doi: 10.1016/S0895-4356(01)00372-9

Blecher, R., Elliott, M. A., Yilmaz, E., Dettori, J. R., Oskouian, R. J., Patel, A., ... Chapman, J. R. (2019). EBSJ Special Section: Systematic Review Contact Sports as a Risk Factor for Amyotrophic Lateral Sclerosis: A Systematic Review. *Global Spine Journal*, *9*(1), 104–118. Retrieved from `https://us.sagepub.com/en-us/nam/open-access-at-sage` doi: 10.1177/2192568218813916

Box, G. (1979). Robustness in Statistics. In *Robustness in the strategy of scientific model building* (pp. 201–236). Elsevier. doi: 10.1016/b978-0-12-438150-6.50018-2

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). *Journal of the neurological sciences*, *169*(1-2), 13–21. doi: 10.1016/s0022-510x(99)00210-5

Cornell Statistical Consulting Unit. (2012). *Interpreting Coefficients in Regression with Log-Transformed Variables.*

Cox, P. A., Kostrzewa, R. M., & Guillemin, G. J. (2018). BMAA and Neurodegenerative Illness. *Neurotoxicity Research*, *33*(1), 178–183. doi: 10.1007/s12640-017-9753-6

David A. Belsley, R. E. W., Edwin Kuh. (2004). Deteting Influential Observations and Outliers. In Wiley (Ed.), *Regression diagnostics: Identifying influential data and sources of collinearity* (chap. 2).

Day, N., & Byar, D. (1979). Testing Hypotheses in Case-Control Studies-Equivalence of Mantel-Haenszel Statistics and Logit Score Tests. , *35*(3), 623–630. Retrieved from `https://www.jstor.org/stable/2530253`

DeCarvalho, M., Ryczkowski, A., Andersen, P., Gromicho, M., Grosskreutz, J., Kuźma-Kozakiewicz, M., ... Miltenberger Miltenyi, G. (2017). International Survey of ALS Experts about Critical Questions for Assessing Patients with ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, *18*(7-8), 505–510. doi: 10.1080/21678421.2017.1349150

Frank E. Harrell, J. (2015). General Aspects of Fitting Regression Models. In *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis* (2nd ed., pp. 21–23). London. doi: 10.1177/096228020401300512

Gauthier, J., Wu, Q. V., & Gooley, T. A. (2019). Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. *Bone Marrow Transplantation*, *55*(4), 675–680. doi: 10.1038/s41409-019-0679-x

Gautier, G., Verschueren, A., Monnier, A., Attarian, S., Salort-Campana, E., & Pouget, J. (2010). ALS with respiratory onset: Clinical features and effects of non-invasive ventilation on the prognosis. *Amyotrophic Lateral Sclerosis*, *11*(4), 379–382. doi: 10.3109/17482960903426543

Hardiman, O., van den Berg, L. H., & Kiernan, M. C. (2011). Clinical diagnosis and management of amyotrophic lateral sclerosis. *Nature Reviews Neurology*, *7*(11), 639–649. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/21989247http://www.nature.com/articles/nrneurol.2011.153` doi: 10.1038/nrneurol.2011.153

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, *16*(9), 965–980. doi: 10.1002/(SICI)1097-0258(19970515)16:9⟨965::AID-SIM509⟩3.0.CO;2-O

IAUL & Department of Statistics. (2002). *The Analysis of Variance.* Retrieved from `http://www.stats.ox.ac.uk/pub/bdr/IAUL/ModellingLecture4.pdf`

Ingre, C., Ross, P., Phiel, F., Kamel, F., & F, F. (2015). Risk factors for amyotrophic lateral sclerosis. *Clinical Epidemiology*, *7*, 181–193. doi: 10.2147/CLEP.S37505

Ito, P. K. (1980). 7 Robustness of ANOVA and MANOVA test procedures. *Handbook of Statistics*, *1*, 199–236. doi: 10.1016/S0169-7161(80)01009-7

Iyer, R., Hosmer, D. W., & Lemeshow, S. (1991). *Applied Logistic Regression.* (Vol. 40) (No. 4). doi: 10.2307/2348743

Kimura, F., Fujimura, C., Ishida, S., Nakajima, H., Furutama, D., Uehara, H., … Hanafusa, T. (2006). Progression rate of ALSFRS-R at time of diagnosis predicts survival time in ALS. *Neurology*, *66*(2), 265–267. doi: 10.1212/01.wnl.0000194316.91908.8a

Korn, E. L., & Graubard, B. I. (2011). Appendix C: Restricted Cubic Regression Splines. In *Analysis of health surveys* (pp. 345–346). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/9781118032619.app3

Lemeshow, S., & David, W. (2014). *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods. 21-Logistic Regression* (N. Balakri ed.; I. P. John Wiley & Sons, Ed.).

Logroscino, G., Traynor, B. J., Hardiman, O., Chiò, A., Mitchell, D., Swingler, R. J., … EURALS (2010). Incidence of amyotrophic lateral sclerosis in Europe. *Journal of neurology, neurosurgery, and psychiatry*, *81*(4), 385–390. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/19710046http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2850819` doi: 10.1136/jnnp.2009.183525

Longinetti, E., & Fang, F. (2019). Epidemiology of amyotrophic lateral sclerosis: An update of recent literature. *Current Opinion in Neurology*, *32*(5), 771–776. doi: 10.1097/WCO.0000000000000730

Nance, R., Delaney, J., McEvoy, J. W., Blaha, M. J., Burke, G. L., Navas-Acien, A., … McClelland, R. L. (2017). Smoking intensity (pack/day) is a better measure than pack-years or smoking status for modeling cardiovascular disease outcomes. *Journal of clinical epidemiology*, *81*, 111–119. doi: 10.1016/j.jclinepi.2016.09.010

National Institute of Neurological Disorders and Stroke NIH. (n.d.). *Amyotrophic Lateral Sclerosis (ALS) Fact Sheet — National Institute of Neurological Disorders and Stroke.* Retrieved from `https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Amyotrophic-Lateral-Sclerosis-ALS-Fact-Sheet`

Paganoni, S., Macklin, E. A., Lee, A., Murphy, A., Chang, J., Zipf, A., … Atassi, N. (2014). Diagnostic timelines and delays in diagnosing amyotrophic lateral sclerosis (ALS). *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, *15*(5-6), 453–456. doi: 10.3109/21678421.2014.903974

Pires, S., Gromicho, M., Pinto, S., Carvalho, M., & Madeira, S. C. (2019). Predicting non-invasive ventilation in ALS patients using stratified disease progression groups. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2018-Novem*, 748–757. doi: 10.1109/ICDMW.2018.00113

Bibliography

Pupillo, E., Messina, P., Giussani, G., Logroscino, G., Zoccolella, S., Chiò, A., … Vertué, G. (2014). Physical activity and amyotrophic lateral sclerosis: A European population-based case-control study. *Annals of Neurology*, *75*(5), 708–716. doi: 10.1002/ana.24150

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., … Traynor, B. J. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, *72*(2), 257–268. doi: 10.1016/j.neuron.2011.09.010

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, *83*(402), 426–431. doi: 10.1080/01621459.1988.10478613

Sullivan, K. M. (n.d.). Two by Two Tables Containing Counts (TwobyTwo).

Wang, H., Reilly, J. O., Weisskopf, M. G., Kolonel, L. N., & Ascherio, A. (2012). Smoking and risk of amyotrophic lateral sclerosis: a pooled analysis of five prospective cohorts. *Arch Neurol*, *68*(2), 207–213. doi: 10.1001/archneurol.2010.367.Smoking

Questionnaire

This annex presents the questionnaire used to assess the functional status of patients with ALS.

ALS Functional Rating Scale Revised (ALS-FRS-R)

Date:………………………………….Name patient:……………………………………………Date of Birth:…………………………………

Patient's number…………………………………………………………………………………..Right-/left-handed

Item 1: SPEECH

4 □ Normal speech process

3 □ Detectable speech disturbance

2 □ Intelligible with repeating

1 □ Speech combined with non-vocal communication

0 □ Loss of useful speech

Item 2: SALIVATION

4 □ Normal

3 □ Slight but definite excess of saliva in mouth; may have nighttime drooling

2 □ Moderately excessive saliva; may have minimal drooling (during the day)

1 □ Marked excess of saliva with some drooling

0 □ Marked drooling; requires constant tissue or handkerchief

Item 3: SWALLOWING

4 □ Normal eating habits

3 □ Early eating problems – occasional choking

2 □ Dietary consistency changes

1 □ Needs supplement tube feeding

0 □ NPO (exclusively parenteral or enteral feeding)

Item 4: HANDWRITING

4 □ Normal

3 □ Slow or sloppy: all words are legible

2 □ Not all words are legible

1 □ Able to grip pen, but unable to write

0 □ Unable to grip pen

Item 5a: CUTTING FOOD AND HANDLING UTENSILS
Patients without gastrostomy → Use 5b if >50% is through g-tube

4 □ Normal

3 □ Somewhat slow and clumsy, but no help needed

2 □ Can cut most foods (>50%), although slow and clumsy; some help needed

1 □ Food must be cut by someone, but can still feed slowly

0 □ Needs to be fed

Item 5b: CUTTING FOOD AND HANDLING UTENSILS
Patients with gastrostomy → 5b option is used if the patient has a gastrostomy and only if it is the primary method (more than 50%) of eating .

4 □ Normal

3 □ Clumsy, but able to perform all manipulations independently

2 □ Some help needed with closures and fasteners

1 □ Provides minimal assistance to caregiver

0 □ Unable to perform any aspect of task

Item 6: DRESSING AND HYGIENE

4 □ Normal function

3 □ Independent and complete self-care with effort or decreased efficiency

2 □ Intermittent assistance or substitute methods

1 □ Needs attendant for self-care

0 □ Total dependence

Item 7: TURNING IN BED AND ADJUSTING BED CLOTHES

4 □ Normal function

3 □ Somewhat slow and clumsy, but no help needed

2 □ Can turn alone, or adjust sheets, but with great difficulty

1 □ Can initiate, but not turn or adjust sheets alone

0 □ Helpless

Item 8: WALKING

4 □ Normal

3 □ Early ambulation difficulties

2 □ Walks with assistance

1 □ Non-ambulatory functional movement

0 □ No purposeful leg movement

Item 9: CLIMBING STAIRS

4 □ Normal

3 □ Slow

2 □ Mild unsteadiness or fatigue

1 □ Needs assistance

0 □ Cannot do

Item 10: DYSPNEA

4 □ None

3 □ Occurs when walking

2 □ Occurs with one or more of the following: eating, bathing, dressing (ADL)

1 □ Occurs at rest: difficulty breathing when either sitting or lying

0 □ Significant difficulty: considering using mechanical respiratory support

Item 11: ORTHOPNEA

4 □ None

3 □ Some difficulty sleeping at night due to shortness of breath, does not routinely use more than two pillows

2 □ Needs extra pillows in order to sleep (more than two)

1 □ Can only sleep sitting up

0 □ Unable to sleep without mechanical assistance

Item 12: RESPIRATORY INSUFFICIENCY

4 □ None

3 □ Intermittent use of BiPAP

2 □ Continuous use of BiPAP during the night

1 □ Continuous use of BiPAP during day & night

0 □ Invasive mechanical ventilation by intubation or tracheostomy

Interviewer's name………

OnWebDuals

This appendix presents all the variables present in the OnWebDuals project.

Table 1. Summary of the clinical results expressed by hierarchy of the most relevant questions.

Main topics	The most important questions – 1st hierarchy	Less important questions – 2nd hierarchy	The least important questions – 3rd hierarchy	Important questions = 1st and 2nd hierarchy
Demographic Features				
	Date of birth Gender Date of diagnosis Date of disease onset	Ethnicity	Local of birth Local of parents birth	none
Disease Features				
	Region of onset Predominant UMN vs LMN manifestations at onset Predominant UMN or LMN signs in pts presenting with dysarthria Predominant proximal versus distal weakness in limb onset pts Respiratory symptoms at onset Disinhibition at onset Language deficit at onset	Drop neck at onset Cramps at onset Predominant side at onset Grammatical errors at onset Delusions at onset	Pain at onset	Fasciculation at onset Weight loss at onset Emotional lability at onset Axial weakness at onset Apathy at onset
Clinical Signs				
	Limb UMN signs (hyperreflexia, spasticity, Babinski sign) Bulbar UMN signs (tongue spasticity, brisk jaw jerk) Limb LMN signs (weakness, fasciculations, atrophy) Bulbar LMN signs (tongue fasciculations, tongue atrophy) Emotional incontinence fasciculations Neck weakness Resting respiratory fatigue Orthopnoea Weak cough Cognition - disinhibition	Limb UMN signs (Hoffman sign, loss of hand dexterity with normal strength) Bulbar UMN sign (jaw clonus, increased retching, brisk pharyngeal reflex) Bulbar LMN signs (facial fasciculations, masseter atrophy, weak orbicularis oris) Paraspinal muscles atrophy Exertional fatigue to minor efforts Extrapyramidal signs Sensory changes Apraxia Postural instability Cognition – grammar errors Cognition – delusions	Blood pressure	Limb LMN signs (hyporeflexia) Thoracic muscles fasciculations Paradoxical respiration Cognition - apathy Cognition - language deficit Cognition – depression Body-mass index

(continued)

Main topics	The most important questions – 1st hierarchy	Less important questions – 2nd hierarchy	The least important questions – 3rd hierarchy	Important questions = 1st and 2nd hierarchy
Disease Severity and Progression Rate	ALSFRS-R at entry ALSFRS-R subscores at entry Pattern of spreading Timing of spreading	Staging		
Investigations	Electromyography CK MRI (brain, spinal cord) Forced Vital Capacity Slow Vital Capacity Maximal Inspiratory Pressure	Albumin Creatinine Chloride Cholesterol Triglycerides Arterial PO2 Arterial PCO2 Nocturnal oximetry Sleep studies		Maximal Expiratory Pressure SNIP
Comorbidities	Diabetes Hypercholesterolaemia Hypertriglyceridaemia Smoking	Autoimmune intestinal disorder Heart – ischaemia Heart – arrhythmia Heart – insufficiency		Blood hypertension Hyperthyroidism Hypothyroidism Autoimmune rheumatologic disorder Stroke Cancer
Previous Medication	Riluzole	Psychiatric drugs Supplements		
Genetic Background	Familial history of ALS - 1st degree Familial history of ALS - 2nd degree	Familial history of ALS - 3rd degree Familial history of Parkinsons disease Parents diseases Parents' diseases and cause of death Siblings diseases and cause of death		Familial history of Alzheimers disease
Habits, previous trauma and surgery	Physical exercise Brain trauma Spine surgery	Other trauma Limb surgery		Abdominal surgery Pelvic Surgery Other surgery
Occupations	Current main occupation Previous occupation	2nd most important occupation		

ALS: amyotrophic lateral sclerosis; LMN: lower motor neuron; UMN: upper motor neuron; pts: patients.
Differences between 1st vs. 2nd vs 3rd degree at hierarchical levels are set $p < 0.05$ (see Methods).

R code

In this thesis, the usual packages for the chosen methodologies were used, the code can be consulted at https://github.com/aRitaH/Thesis/tree/master

We chose to present only the code developed by us to create the 3.1 and 4.12 plots and the Stuckel function, which, is not from our authorship, but is important to share due to the difficulty of access.

```
   # Figure 3.1

   GRa<-read_excel("/Users/ritahenriquesepidoc/Desktop/Tese/GRa.xlsx")
   head(GRa)
   data<-GRa
   dim(data)
   names(data)<-c("inicio","fim","inicio1", "fim1")
   data<-data[complete.cases(data),]
   data<-data[order(data$fim),]
   n<-nrow(data)

   li<-min(data$inicio);li
   ls<-max(data$fim);ls
   plot(c(li,ls),c(1,n),type="n",ylab = "ID",xlab = "Smoking Days")
   for(i in 1:n)
   segments(data$inicio[i],i,data$fim[i],i,col="grey")
   abline(v=0,col=2)

   ## In Years

   data<-data[complete.cases(data),]
   data<-data[order(data$fim1),]
   n<-nrow(data)
   lia<-min(data$inicio1);lia
   lsa<-max(data$fim1);lsa
   plot(c(lia,lsa),c(1,n),type="n",ylab = "ID",xlab = "Smoking Years"))
,cex.main=2, cex.lab=2, cex.axis=2,cex.id = 1.5,cex.caption =2 )
   axis(side=1, at=seq(-60, 10, by=10))
   for(i in 1:n) segments(data$inicio1[i],i,data$fim1[i],i,col="grey")
```

. R CODE

```
   abline(v=0,col=2)

   # Function for Stukel test

   stukel <- function(object, alternative = c("both", "alpha1", "alpha2"))

   DNAME <- deparse(substitute(object))
   METHOD <- "Stukel’s test of the logistic link"
   alternative <- match.arg(alternative)
   eta <- predict(object, type = "link")
   etasq <- 0.5 * eta * eta
   is positive?
   etapos <- eta > 0
   dv <- matrix(0, nrow = length(eta), ncol = 2)
   dv[etapos, 1] <- etasq[etapos]
   dv[!etapos, 2] <-etasq[!etapos]
   colnames(dv) <- c("z1", "z2")
   oinfo <- stats::vcov(object)
   ### qr decomposition of matrix
   oX <- qr.X(object$qr)
   ImH <- - oX %*% oinfo %*% t(oX)
   diag(ImH) <- 1 + diag(ImH)
   wdv <- sqrt(object$weights) * dv
   qmat <- t(wdv) %*% ImH %*% wdv
   sc <- apply(dv * (object$weights * residuals(object, "working")), 2,
sum)
   allstat <- c(sc * sc / diag(qmat), sc %*% solve(qmat) %*% sc)
   names(allstat) <- c("alpha1", "alpha2", "both")
   allpar <- c(1,1,2)
   names(allpar) <- names(allstat)
   allpval <- pchisq(allstat, allpar, lower.tail=FALSE)
   STATISTIC <- allstat[alternative]
   PARAMETER <- allpar[alternative]
   names(PARAMETER) <- "df"
   PVAL <- allpval[alternative]
   names(allpar) <- rep("df", 3)
   structure(list(statistic = STATISTIC,
   parameter = PARAMETER,
   p.value = PVAL,
   alternative = alternative,
   method = METHOD, data.name = DNAME,
   allstat = allstat, allpar = allpar, allpval = allpval
   ),
   class = "htest")
```

```
## Figure 4.12- Evaluate linearity of logit

#Age
data.ALS<-ALS[order(ALS$Age),]
qnt <- quantile(data.ALS$Age, probs=seq(0, 1, 0.25),na.rm = T)
groups <- cut(data.ALS$Age, breaks=qnt[1:5], include.lowest=TRUE, labels=FALSE)
data.ALS <- cbind(data.ALS, groups)
mid.points <- qnt[-length(qnt)] + diff(qnt)/2
prop.suc <- c()
logit.suc <- c()
for(i in 1:4) group <- subset(data.ALS, data.ALS$groups == i)
prop.suc[i] <-nrow(group[group$Type=="Patients"])/nrow(group)
logit.suc[i] <-log(prop.suc[i]/(1-prop.suc[i]))
par(pty="s")
plot(mid.points, logit.suc, type=’n’, las=1, ylab="logit Age", xlab="midpoint Age", cex.lab=1.2, cex.axis=1.2) points(mid.points, logit.suc, pch=16, cex=1.2, col=1)
abline(lm(logit.suc   mid.points), lty=2)
mod <- lm(logit.suc   mid.points)
coef(mod)
grid(col = "lightgray")
par(pty="m")

#TE
data.ALS<-ALS[order(ALS$TE),]
qnt <- quantile(data.ALS$TE, probs=seq(0, 1, 0.25),na.rm = T)
groups <- cut(data.ALS$TE, breaks=qnt[1:5], include.lowest=TRUE, labels=FALSE)
data.ALS <- cbind(data.ALS, groups)
mid.points <- qnt[-length(qnt)] + diff(qnt)/2
prop.suc <- c()
logit.suc <- c()
for(i in 1:4)
group <- subset(data.ALS, data.ALS$groups == i)
prop.suc[i] <-nrow(group[group$Type=="Patients"])/nrow(group)
logit.suc[i] <-log(prop.suc[i]/(1-prop.suc[i]))
par(pty="s")
plot(mid.points, logit.suc, type=’n’, las=1, ylab="logit TE", xlab="midpoint TE", cex.lab=1.2, cex.axis=1.2)
points(mid.points, logit.suc, pch=16, cex=1.2, col=1)
abline(lm(logit.suc   mid.points), lty=2)
mod <- lm(logit.suc   mid.points)
coef(mod)
grid(col = "lightgray")
```

. R CODE

```
   par(pty="m")

   breaks = agebreaks, right = FALSE, labels = agelabels)]

   group <- subset(data.ALS, data.ALS$groups == i)
   prop.suc[i] <-nrow(group[group$Type=="Patients"])/nrow(group)
   logit.suc[i] <-log(prop.suc[i]/(1-prop.suc[i]))
   type='n', las=1, ylab="logit Age", xlab="midpoint Age", cex.lab=1.2,
cex.axis=1.2)
   #TE
   data.ALS<-ALS[order(ALS$TE),]
   qnt <- quantile(data.ALS$TE, probs=seq(0, 1, 0.25),na.rm = T)
   groups <- cut(data.ALS$TE, breaks=qnt[1:5], include.lowest=TRUE, la-
bels=FALSE)
   data.ALS <- cbind(data.ALS, groups)
   mid.points <- qnt[-length(qnt)] + diff(qnt)/2
   prop.suc <- c()
   logit.suc <- c()
   for(i in 1:4)
   group <- subset(data.ALS, data.ALS$groups == i)
   prop.suc[i] <-nrow(group[group$Type=="Patients"])/nrow(group)
   logit.suc[i] <-log(prop.suc[i]/(1-prop.suc[i]))
   par(pty="s")
   plot(mid.points, logit.suc, type='n', las=1, ylab="logit TE", xlab="midpoint
TE", cex.lab=1.2, cex.axis=1.2)
   points(mid.points, logit.suc, pch=16, cex=1.2, col=1)
   abline(lm(logit.suc   mid.points), lty=2)
   mod <- lm(logit.suc   mid.points)
   coef(mod)
   grid(col = "lightgray")
   par(pty="m")
```