COPPE/UFRJ

TÓPICOS EM APRENDIZADO ESTATÍSTICO NA PESQUISA MÉDICA

Emília Matos do Nascimento

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Produção.

Orientador: Basilio de Bragança Pereira

Rio de Janeiro
Setembro de 2010

TÓPICOS EM APRENDIZADO ESTATÍSTICO NA PESQUISA MÉDICA

Emília Matos do Nascimento

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE PRODUÇÃO.

Examinada por:

Prof. Basilio de Bragança Pereira, Ph.D.

Prof. Helio dos Santos Migon, Ph.D.

Prof. Nelson Spector, D.Sc.

Prof. Marina Silva Paez, D.Sc.

Prof. Aline Araújo Nobre, D.Sc.

Prof. José Manoel de Seixas, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2010

Nascimento, Emília Matos do

Tópicos em Aprendizado Estatístico na Pesquisa Médica/Emília Matos do Nascimento. – Rio de Janeiro: UFRJ/COPPE, 2010.

XI, 201 p.: il.; 29, 7cm.

Orientador: Basilio de Bragança Pereira

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Produção, 2010.

Referências Bibliográficas: p. 57 – 65.

1. Análise de Sobrevivência. 2. Árvores de Decisão. 3. Estimação e Escolha de Modelos. 4. Modelos Lineares Generalizados. 5. Redes Neurais Artificiais. I. Pereira, Basilio de Bragança. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção. III. Título.

Aos meus pais
Manoel R. Nascimento
(in memoriam) e
Tereza M. Nascimento,
pelo carinho e dedicação
de toda a vida.

Agradecimentos

Agradeço a Deus, em primeiro lugar, por ter-me dado a força e a auto-confiança necessárias para perserverar e vencer os meus próprios limites.

Ao Professor Basilio de Bragança Pereira, por quem tenho grande respeito e admiração, não somente pelo seu profissionalismo e competência, mas pelo grande ser humano que é. Agradeço pela sua atenção, apoio, incentivo e por compartilhar os seus conhecimentos, propiciando o meu aprendizado através do seu valioso trabalho de orientação.

Aos Professores Helio dos Santos Migon, Nelson Spector, Marina Silva Paez, Aline Araújo Nobre e José Manoel de Seixas, por terem participado da banca examinadora, prestigiando-me com suas relevantes contribuições.

Ao Dr. Joaquim Ribeiro Filho, Dr. Roberto Coury Pedrosa e a todos os co-autores, cujos artigos compõem os apêndices desta tese, por terem confiado as suas bases de dados, peças fundamentais para o desenvolvimento deste trabalho, dando-me a preciosa oportunidade de participar de suas pesquisas.

Ao corpo docente do Programa de Engenharia de Produção e do Instituto de Matemática, que muito contribuiu para o meu crescimento acadêmico.

Ao meu pai, que apesar da distância, está sempre presente nas minhas lembranças, nas minhas orações. À minha mãe, por tanto amor, estímulo e cuidado. À minha tia Zoraide e à minha irmã Maria de Fátima por comemorarem cada uma das minhas conquistas. Ao meu irmão Guido e ao meu sobrinho Giovani por fazerem parte da minha torcida. À minha irmã Celeste, que muito me ajudou com seus conhecimentos de *designer* gráfico. Ao meu futuro cunhado, Márlon Ribeiro, pela manutenção do meu computador, minha ferramenta de trabalho.

Aos meus amigos, especialmente, à Luzia Tonon, Rubens Oliveira, Valdir Melo e Vânia Marins, por suas importantes colaborações e por terem estado sempre presentes nos momentos em que mais precisei de ajuda.

Aos funcionários da secretaria e, em especial à Andréia Lima S. Moreira, sempre cuidando com eficiência das tarefas administrativas.

A todos que, de alguma forma, contribuiram para a realização deste trabalho.

À CAPES pelo apoio financeiro através da concessão da bolsa de estudos.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

TÓPICOS EM APRENDIZADO ESTATÍSTICO NA PESQUISA MÉDICA

Emília Matos do Nascimento

Setembro/2010

Orientador: Basilio de Bragança Pereira

Programa: Engenharia de Produção

Os avanços computacionais exigem pesquisadores, cada vez mais, aptos a trabalhar com grandes bases de dados, sejam de natureza quantitativa ou qualitativa. A utilização do aprendizado estatístico é imprescindível, sobretudo na análise de dados clínicos. A integração dos estatísticos com os pesquisadores da área médica, bem como os de outros campos do conhecimento, é fundamental para o desenvolvimento das diversas pesquisas, desde a fase da coleta dos dados e desenho do estudo até a interpretação dos resultados. Esta tese mostra algumas aplicações das técnicas de aprendizado estatístico na pesquisa médica, como um trabalho conjunto com pesquisadores clínicos do Hospital Universitário Clementino Fraga Filho (HUCFF / UFRJ), resultando em artigos escritos em parceria com os mesmos. Dentre as técnicas utilizadas encontram-se os Modelos Lineares Generalizados, Análise de Sobrevida, Árvores de Decisão, Análise de Componentes Principais e Redes Neurais Artificiais.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

TOPICS IN STATISTICAL LEARNING IN MEDICAL RESEARCH

Emília Matos do Nascimento

September/2010

Advisor: Basilio de Bragança Pereira

Department: Production Engineering

The computational advances require researchers increasingly able to work with large databases being of quantitative or qualitative nature. The use of statistical learning is essential, especially in the analysis of clinical data. The integration of statisticians with the medical researchers, as well as other fields of knowledge, is fundamental to the development of several researches, from the data collection and study design until interpretation of results. This thesis shows some applications of statistical learning in medical research, as a joint work with clinical researchers at the Hospital Universitário Clementino Fraga Filho (HUCFF / UFRJ) resulting in articles written in partnership with them. Among the techniques used are the Generalized Linear Models, Survival Analysis, Decision Trees, Principal Component Analysis, and Neural Networks.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

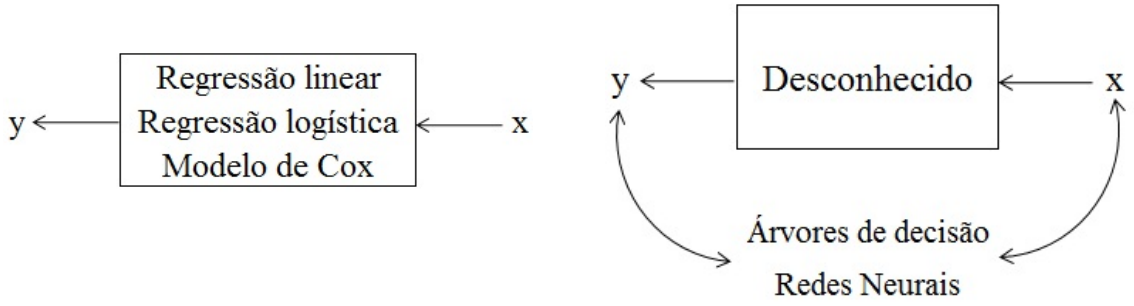
Introdução

No passado, pouco interesse foi demonstrado nas ferramentas estatísticas voltadas para indução e descrição. A busca do modelo "correto" e o seu valor p associado tornou a análise exploratória de dados uma atividade secreta. Os avanços computacionais, da matemática aplicada e da estatística fizeram da análise exploratória um privilégio, que trouxe uma variedade de novas idéias e conceitos, especialmente, na estatística computacional (Berk, 2008). Avanços estes, que exigem que cientistas de diversas áreas do conhecimento sejam, cada vez mais, capazes de trabalhar com bases de dados quantitativos e qualitativos, dispostos em matrizes de grandes dimensões, como por exemplo, na pesquisa clínica (Scully et al., 1997; Prather et al., 1997; Mullins et al., 2006).

Essas novas abordagens recebem diversos nomes, tais como aprendizado estatístico (*statistical learning*), mineração de dados (*data mining*) e máquinas de aprendizagem (*machine learning*) (Berk, 2005; Hastie et al., 2009), com várias aplicações na área médica (Prather et al., 1997; Holmes et al., 2000; Downs & Wallace, 2000). Entretanto, apesar da conexão óbvia entre aprendizado estatístico e Estatística, muitas metodologias usadas em aprendizado estatístico tiveram origem em outros campos (Friedman, 1997).

Na modelagem estatística, Breiman (2001) menciona a existência de duas culturas: a primeira assume que os dados sejam gerados segundo um modelo estocástico (Figura 1.1(a)); e a segunda utiliza modelagem algorítmica (Figura 1.1(b)).

Considerando o aprendizado estatístico, Hastie et al.(2009) descreve um cenário onde se deseja predizer uma saída, geralmente uma variável quantitativa ou categórica, a partir de um conjunto de características (variáveis explicativas ou preditivas). Neste tipo de análise, utiliza-se um conjunto de treinamento dos dados, no qual se observa a saída (desfecho), dado um conjunto de variáveis de entrada; e constrói-se um modelo preditivo, ou de aprendizado, capaz de predizer uma saída (desfecho) para uma nova entrada ainda não vista. Este procedimento é denominado problema



(a) Estatística: Teoria em busca de dados (b) *Data Mining*: Dados em busca de teoria

Figura 1.1: Duas Culturas

de aprendizado supervisionado. No problema de aprendizado não supervisionado, apenas as variáveis explicativas são observadas. Neste caso, onde a variável de saída (desfecho) não é conhecida, analisam-se como os dados estão organizados, ou agrupados.

Em relação à técnica do aprendizado, duas interpretações diferentes são encontradas na literatura: a primeira está fortemente relacionada ao pensamento estatístico tradicional (Lovell, 1983 e Chatfield, 1995; apud Coppi, 2002), onde a maior preocupação está na confiabilidade dos procedimentos de testes de hipóteses e na seleção de modelos; e a segunda diz respeito à pesquisa na área de *Knowledge Discovery in Databases* - KDD (Frawley et al., 1992), cujo objetivo é extrair informações de grandes bases de dados.

Os métodos relacionados ao KDD, aliados à possibilidade da modelagem de sistemas complexos, conduziram ao surgimento de dois paradigmas (Pereira, 2000), ilustrados nas Figuras 1.2 e 1.3:

Paradigma I: "Estatística: Teoria em busca de dados confirmatórios"
Paradigma II: "Aprendizado estatístico: Dados em busca de uma teoria"

A Figura 1.2 mostra o primeiro paradigma, iniciando com a observação do fenômeno; passando pelas etapas do desenho do experimento, com o uso de técnicas de Planejamento de Experimentos ou da Teoria da Amostragem; e apresentando as informações através de Estatísticas descritivas ou por intermédio da análise exploratória dos dados. Os resultados obtidos experimentalmente são, então, comparados aos esperados por meio de testes de hipóteses, quando são utilizadas técnicas de Inferência Estatística, tais como inferência em Séries Temporais, Estatística Multivariada, Estatística Não Paramétrica e Análise de Decisões. A técnica de inferência a ser adotada depende do tipo de dado e do problema em análise. Finalmente, a adequação do modelo é verificada através de técnicas de análise dos resíduos e ajuste do modelo. Assim, o conhecimento adquirido é utilizado para a formulação de no-

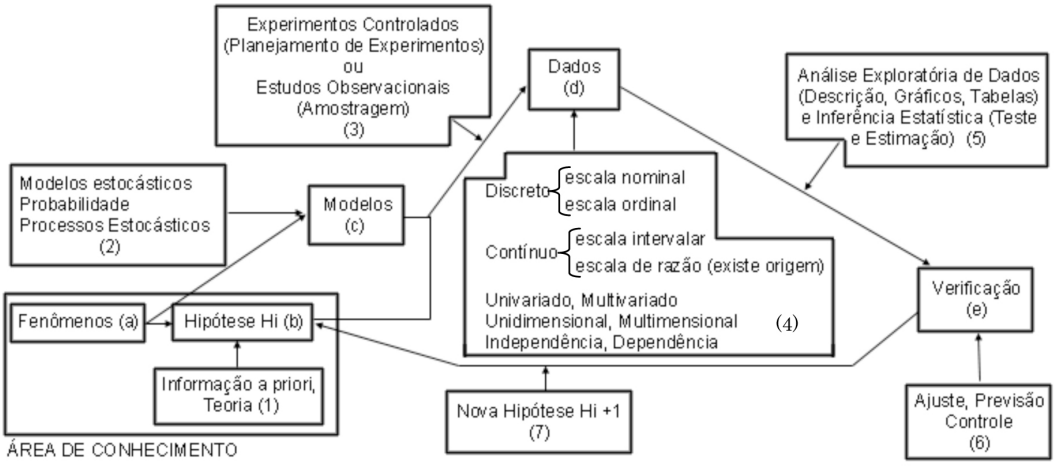


Figura 1.2: Paradigma I - Estatística

vas hipóteses e alteração do modelo, dando início a um novo ciclo (Pereira, 1997; Pereira, 2000).

A Figura 1.3 ilustra o segundo paradigma partindo da base de dados e utilizando métodos tais como Redes Neurais, *Machine Learning* e métodos estatísticos para formulação de hipóteses; fazendo ajustes e previsão através de modelos, a partir da observação dos fenômenos; e, finalmente, fazendo a verificação da adequação do modelo. Procedimentos esses, que permitem identificar padrões ocultos nos dados e transformá-los em conhecimento (Pereira, 1997; Pereira, 2000).

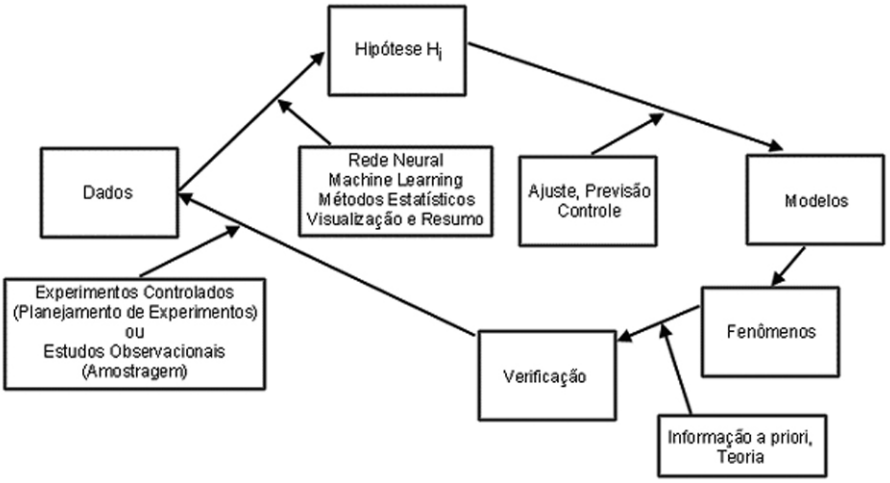


Figura 1.3: Paradigma II - *Data Mining*

Esses dois paradigmas devem, assim, ser utilizados para a obtenção do conhecimento expresso pela equação de Rao (Rao, 1997; apud Pereira, 2000), mostrada na Figura 1.4.

Segundo Coppi (2002), a informação a ser obtida pode ser constituída de regras que descrevem as propriedades dos dados (associações ou relacionamentos entre variáveis), padrões, *clusters* de objetos contidos na base de dados, entre outras.

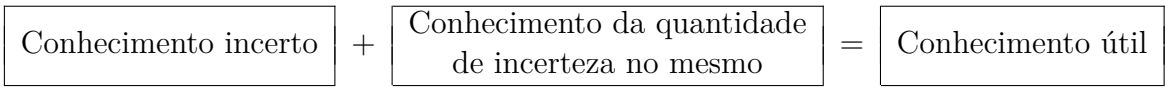


Figura 1.4: Equação de Rao

Deste modo, os sistemas que utilizam *Data Mining*, podem ser vistos como casos particulares de sistemas de aprendizagem ou cognitivos, quando aplicados a uma base de dados. Em seu artigo, Coppi (2002) propõe uma abordagem que combina as características de ambas as interpretações, enfatizando a falta de um quadro teórico e justificando a prática do aprendizado estatístico.

É comum assumir, erroneamente, que a presença do estatístico seja necessária apenas para análise dos resultados. O estatístico desempenha um papel relevante na análise dos dados, contudo a sua participação é importante mesmo antes da fase inicial da pesquisa, ou seja, desde o seu planejamento (Pocock, 1983; apud Martinez, 2009). Assim, a colaboração entre pesquisadores clínicos e estatísticos é fundamental para a realização de um ensaio clínico bem sucedido. Embora o desenho do estudo e as análises estatísticas sejam de principal responsabilidade do estatístico, a compreensão dos conceitos básicos da estatística é essencial para os pesquisadores clínicos envolvidos no estudo (Green et al., 2003). Altman & Bland (1991) abordaram a necessidade do conhecimento estatístico na pesquisa médica, ressaltando a importância da compreensão das idéias estatísticas pelos pesquisadores clínicos. Os estatísticos, por sua vez, devem também compreender os aspectos clínicos (Pereira, 2000).

Este trabalho tem como objetivo mostrar a aplicabilidade das técnicas de aprendizado estatístico na pesquisa médica, visando à integração entre pesquisadores da área clínica e de estatística. A relevância deste trabalho se dá na medida em que possibilite a identificação dos fatores prognósticos clínicos relacionados a cada estudo, auxiliando no tratamento e no diagnóstico.

Algumas aplicações do aprendizado estatístico são apresentadas, nesta tese, resultando em artigos escritos em parceria com pesquisadores do Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ), alguns submetidos e outros já publicados ou aceitos para publicação.

O trabalho de implementação foi realizado, em sua maioria, com o uso do "software" R (R Development Core Team, 2009).

1.1 Organização do Trabalho

O Capítulo 2 faz uma descrição das técnicas de aprendizado estatístico, utilizadas durante a realização desta tese. A discussão e as conclusões são apresentadas no Capítulo 3.

Os Apêndices apresentam os artigos publicados ou submetidos, que são o resultado de consultorias realizadas na Divisão de Pesquisa do - HUCFF/UFRJ, onde diferentes métodos estatísticos foram aplicados:

B. Árvores de sobrevida e modelo de riscos proporcionais de Cox;
C. Estimador de Kaplan-Meier e comparação de curvas de sobrevida;
D. Algoritmo EM (*Expectation-Maximization*) e Critério de Informação de Akaike (AIC);
E. Redes neurais do tipo *feedforward*;
F. Modelos log-lineares;
G. Uso não-convencional das curvas de sobrevida, para comparação de dados qualitativos ordinais;
H. Árvores de regressão e curvas de sobrevida para análise de concordância entre medidas (*Survival Agreement Plot*); e
I. Revisão de estudos e metodologias (incluindo as redes neurais do tipo *feedforward*) para associação entre poluição atmosférica doenças respiratórias.

Embora não relacionado à área médica, o Apêndice A apresenta, para efeito ilustrativo, o uso combinado do aprendizado supervisionado e não-supervisionado através das redes neurais do tipo SOM (*Self Organizing Map*), para clusterização, e MLP (*Multilayer Perceptron*), para identificação de variáveis relevantes.

Capítulo 2

Algumas Técnicas de Aprendizado Estatístico

Neste capítulo, serão apresentadas algumas técnicas de aprendizado estatístico, mais especificamente, as utilizadas na elaboração dos artigos mencionados na Seção 1.1, escritos em parceria com pesquisadores do HUCFF-UFRJ.

2.1 Modelos Lineares Generalizados

Os modelos clássicos lineares e de mínimos quadrados tiveram início no trabalho de Gauss & Legendre (Stigler, 1981 e Stigler, 1986; apud McCullagh & Nelder, 1989), para estudo de dados astronômicos, onde a distribuição Normal ou Gaussiana foi utilizada para descrever as propriedades dos erros de medição. No século XIX, os modelos lineares foram amplamente utilizados em aplicações biológicas, assumindo-se também a normalidade dos erros (McCullagh & Nelder, 1989).

Gauss mostrou que muitas propriedades importantes das estimativas de mínimos quadrados não dependiam da normalidade, e sim, da suposição de variância constante e de independência. Uma propriedade semelhante pode ser aplicada aos modelos lineares generalizados. Ou seja, embora sejam feitas referências a várias distribuições padrões, tais como a Normal, binomial, Poisson, exponencial ou gama, as propriedades de segunda ordem (esperança dos erros igual a zero e variância constante) dos parâmetros não são sensíveis à forma de distribuição assumida (McCullagh & Nelder, 1989).

Na notação matricial, os modelos clássicos lineares utilizam um vetor de observações $y = \{y_1, \ldots, y_n\}^T$, representando a variável resposta; um conjunto de variáveis explicativas (covariáveis) expresso por uma matriz $\boldsymbol{X}$, de ordem $n \times p$, onde cada linha refere-se a uma observação e cada coluna, a uma covariável; e, associado às covariáveis, um vetor de parâmetros $\beta = \{\beta_1, \ldots, \beta_p\}^T$ a serem estimados. Nessas

condições, define-se, para cada valor de β, um vetor de resíduos $\epsilon(\beta) = y - \boldsymbol{X}\beta$.

Os Modelos Lineares Generalizados (GLM) (Nelder & Wedderburn, 1972) constituem uma extensão dos modelos de regressão linear, admitindo-se uma distribuição não-normal para a variável resposta e permitindo a modelagem como uma função da média. Os GLM são formados por três componentes (Agresti, 2002; Agresti, 2007):

1. Um componente aleatório associando a variável resposta ($\boldsymbol{Y}$) à sua distribuição de probabilidade. A variável dependente $\boldsymbol{Y}$ é composta por n observações independentes, que seguem uma distribuição da família exponencial, definida como (McCullagh & Nelder, 1989)

$$f_Y(y; \theta, \phi) = exp[(y\theta - b(\theta))/a(\phi) + c(y, \phi)],$$

onde θ é um parâmetro e $a(.)$, $b(.)$ e $c(.)$ são funções específicas. A Tabela 2.1 apresenta os identificadores de algumas distribuições que fazem parte da família exponencial (Demétrio, 2002). Os modelos Normal, Gama e Normal Inversa são utilizados para variáveis contínuas. O modelo de Poisson é utilizado para dados de contagem e o binomial, para proporções.

2. Uma componente sistemática, também chamada preditor linear, que descreve as variáveis explicativas como uma função linear

$$\eta_i = \sum_j \beta_i x_{ij}, \qquad i = 1, \ldots, n; \qquad j = 1, \ldots, p.$$

3. Uma função de ligação $g(.)$ que associa $\mu = E(\boldsymbol{Y})$, valor esperado de $\boldsymbol{Y}$, à componente sistemática

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, \ldots, n.$$

A função de ligação identidade é a mais simples e especifica um modelo linear para a resposta média, $\mu = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p$, sendo utilizada nos modelos de regressão para respostas contínuas (Agresti, 2007).

Em geral, as seguintes funções de ligação são usadas (Agresti, 2007; McCullagh & Nelder, 1989):

1. Logito ou função logística: $g_1(p) = \log(p/(1-p))$, que modela o logaritmo da razão de chances, linearizando a variável preditora;

2. Probito ou função Normal inversa: $g_2(p) = \Phi^{-1}(p)$ onde Φ^{-1} é a função de distribuição normal inversa cumulativa;

3. Complementar log-log: $g_3(p) = \log(-\log(1-p))$; e

4. Log-log: $g_4(p) = -\log\Big(-\log(p)\Big)$, relacionada à função complementar log-log. A função log-log é raramente utilizada, por não ser apropriada para os casos onde $p < 0,5$.

As quatro funções podem ser obtidas através da inversa de funções de distribuição acumuladas conhecidas, definidas em $\Re$. As duas primeiras funções são simétricas, considerando que $g_1(p) = -g_1(1-p)$. As duas últimas, apesar de não simétricas, são relacionadas entre si já que $g_3(p) = -g_4(1-p)$ (McCullagh & Nelder, 1989).

McCullagh & Nelder (1989) mostraram, graficamente, que: (1) as funções logito e probito são quase linearmente relacionadas no intervalo $0,1 \leq p \leq 0,9$; (2) para valores pequenos de p, a função complementar log-log e a função logito aproximam-se de $\log(p)$ e; (3) quando p aproxima-se de 1, tanto a função log-log quanto a complementar log-log aproximam-se do infinito, muito mais lentamente do que a função logística ou a probito. As origens e a evolução dos modelos logito e probito são apresentadas por Cramer (2003).

Os modelos lineares generalizados incluem os modelos de regressão linear e análise de variância; os modelos logito e probito, para respostas binárias; os modelos log-lineares; e os modelos multinomiais, para dados de contagem (McCullagh & Nelder, 1989). A Tabela 2.2 indica os métodos mais utilizados na análise estatística para os diferentes tipos de variáveis resposta e variáveis explicativas (Dobson, 2002).

As seções 2.1.1, 2.1.2 , 2.1.3 e 2.1.4 apresentam, respectivamente, os modelos logístico, multinomiais, Log-lineares e de Poisson.

Tabela 2.1: Identificadores de distribuições da Família Exponencial

	$a(\phi)$	θ	$b(\theta)$	$c(y;\phi)$	$\mu(\theta)$	Variância
Normal $N(\mu,\sigma^2)$	σ^2	μ	$\dfrac{\theta^2}{2}$	$-\dfrac{1}{2}\left[\dfrac{y^2}{\sigma^2}+\ln(2\pi\sigma^2)\right]$	θ	1
Poisson $P(\mu)$	1	$\ln\ \mu$	e^θ	$-\ln\ y!$	e^θ	μ
Binomial $B(m,\pi)$	1	$\ln\left(\dfrac{\pi}{1-\pi}\right)$	$m\ \ln(1+e^\theta)$	$\ln\ \binom{m}{y}$	$m\left(\dfrac{e^\theta}{1+e^\theta}\right)$	$\dfrac{1}{m}\big[\mu(m-\mu)\big]$
Binomial Negativa $BN(\mu,k)$	1	$\ln\left(\dfrac{\mu}{\mu+k}\right)$	$-k\ \ln(1-e^\theta)$	$\ln\left[\dfrac{\Gamma(k+y)}{\Gamma(k)\,y!}\right]$	$k\left(\dfrac{e^\theta}{1-e^\theta}\right)$	$\mu\left(\dfrac{\mu}{k}+1\right)$
Gama $G(\mu,\nu)$	ν^{-1}	$-\dfrac{1}{\mu}$	$-\ln(-\theta)$	$\nu\ \ln(\nu y)-\ln\ y-\ln\ \Gamma(\nu)$	$-\dfrac{1}{\theta}$	μ^2
Normal Inversa $IG(\mu,\sigma^2)$	σ^2	$-\dfrac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	$-\dfrac{1}{2}\left[\ln(2\pi\sigma^2y^3)+\dfrac{1}{\sigma^2y}\right]$	$(-2\theta)^{-1/2}$	μ^3

Tabela 2.2: Métodos Estatísticos mais utilizados para os diferentes tipos de variáveis resposta e variáveis explicativas

Variável Dependente	Variáveis Explicativas	Método Estatístico
Contínua	Binária	Teste t
	Nominais (mais que duas categorias)	Análise de Variância
	Ordinais	Análise de Variância
	Contínuas	Regressão Múltipla
	Nominais e Contínuas	Análise de Covariância
	Categóricas e Contínuas	Regressão Múltipla
Binária	Categóricas	Tabelas de Contingência e Regressão Logística
	Contínuas	Modelos Logístico, Probito e outros dose-resposta
	Categóricas e Contínuas	Regressão Logística Nominal
Nominal (mais que duas categorias)	Nominais	Tabelas de Contingência
	Categóricas e Contínuas	Regressão Logística Nominal
Ordinal	Categóricas e Contínuas	Regressão Logística Ordinal
Contagem	Categóricas	Modelo Log-linear
	Categóricas e Contínuas	Regressão de Poisson
Tempos de falha	Categóricas e Contínuas	Análise de Sobrevivência
Respostas correlacionadas	Categóricas e Contínuas	Equações de Estimação Generalizadas e Modelagem Multinível

2.1.1 Modelo Logístico

Os modelos de regressão logística são utilizados quando a variável resposta $\boldsymbol{Y}$ é qualitativa binária (0 ou 1). Nestes modelos, o parâmetro p de interesse está relacionado à proporção de indivíduos, cuja característica está associada ao valor 1.

A distribuição da variável $\boldsymbol{Y}$ terá média e variância, respectivamente, expressas por $E(Y) = 1(p)+0(1-p) = p$ e $Var(Y) = E(Y^2)-(E(Y)^2) = 1^2(p)+0^2(1-p)-p^2 = p(1-p)$. Entretanto, a variância não será constante e estará mais próxima de zero à medida que p se aproxima de 0 ou 1. Assim, a probabilidade p pode ser definida como

$$p(x) = E(Y|X) = \beta_0 + \beta_1 x + \epsilon.$$

Este modelo teria, porém, sérias implicações (Johnson & Wichern, 2007): (1) o valor previsto para Y poderia ser maior que 1 ou menor que zero, pois a expressão linear para o seu valor esperado é ilimitado; e (2) como mencionado anteriormente, a suposição de variância constante poderia ser violada. Assim, em vez de se utilizar o modelo linear para modelar a probabilidade p, considera-se inicialmente a razão de chances (*odds-ratio*), $p/(1-p)$, que é definida como a razão entre a proporção dos indivíduos codificados com o valor 1 e a dos codificados com o valor zero. É importante observar, entretanto, que a razão de chances pode assumir um valor superior a 1.

A implementação do modelo logístico se dá através da modelagem do logaritmo neperiano da razão de chances, que é uma função da probabilidade p. No modelo mais simples, esta função promove a linearização da variável preditora $\boldsymbol{X}$, obtendo-se, então,

$$logit(p) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x. \tag{2.1}$$

A probabilidade p pode ser obtida a partir da exponenciação da equação 2.1, obtendo-se

$$p(y|x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + exp(-\beta_0 - \beta_1 x)}\,.$$

2.1.2 Modelos Multinomiais

Os modelos multinomiais constituem uma extensão dos modelos logísticos, sendo usados nos casos em que a variável resposta é nominal com mais de duas categorias. Caso essas categorias possuam uma ordenação, o modelo de regressão logística ordinal deverá ser utilizado.

2.1.2.1 Modelos Logitos Multinomiais

Para variáveis resposta nominais, utilizam-se os logitos multinomiais combinando-se cada categoria com uma categoria base, escolhida arbitrariamente. Seja J o número de categorias para a variável resposta nominal Y, que segue uma distribuição multinomial. Seja, ainda, J a categoria base. Nestas condições, o logito da categoria base com um preditor x é expresso como

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \qquad j = 1, \ldots, J-1.$$

O modelo possui $J-1$ equações e os efeitos variam de acordo com a categoria. Quando $J = 2$, o modelo reduz-se a uma única equação, resultando no modelo de regressão logístico ordinário para resposta binária (Agresti, 2007)

$$\log\left(\frac{\pi_j}{\pi_J}\right) = logit(\pi_1). \tag{2.2}$$

Por exemplo, para as categorias a e b e considerando J a categoria de base, o modelo 2.2 assume a forma

$$\begin{aligned}
\log\left(\frac{\pi_a}{\pi_b}\right) &= \log\left(\frac{\pi_a/\pi_J}{\pi_b/\pi_J}\right) \\
&= \log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right) \\
&= (\alpha_a + \beta_a x) - (\alpha_b + \beta_b x) \\
&= (\alpha_a - \alpha_b) + (\beta_a - \beta_b)x,
\end{aligned}$$

que equivale ao modelo $\alpha + \beta x$, de parâmetros $\alpha = \alpha_a - \alpha_b$ e $\beta = \beta_a - \beta_b$. As probabilidades de resposta podem ser estimadas pelo modelo (Agresti, 2007)

$$\pi_j = \frac{\exp(\alpha_j + \beta_j x)}{\sum_h \exp(\alpha_h + \beta_h x)}, \qquad j = 1, \ldots, J.$$

2.1.2.2 Modelo de Regressão Logística Ordinal

O modelo de regressão logística ordinal é utilizado quando as categorias de uma variável dependente nominal possui algum tipo de ordenação. Neste caso, utiliza-se a probabilidade cumulativa, que representa a probabilidade de se obter uma resposta até um determinado valor. Assim, para a categoria j, essa probabilidade é expressa por $P(Y \leq j) = \pi_1 + \ldots + \pi_j, \quad j = 1, \ldots, J$ (Agresti, 2007).

As probabilidades cumulativas consideram a ordem existente nas categorias, de forma tal que $P(Y \leq 1) \leq P(Y \leq 2) \leq \ldots \leq P(Y \leq J) = 1$. Os logitos

cumulativos, logitos das probablidades cumulativas, são definidos por (Agresti, 2007)

$$\begin{aligned} logit\Big[P(Y \leq j)\Big] &= \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) \\ &= \log\left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right), \qquad j = 1, \ldots, J-1. \end{aligned}$$

Cada logito cumulativo utiliza todas as categorias da variável desfecho. Assim, para $J = 3$, obtém-se

$$\begin{aligned} logit\Big[P(Y \leq 1)\Big] &= \log\left(\frac{\pi_1}{\pi_2 + \pi_3}\right), \\ logit\Big[P(Y \leq 2)\Big] &= \log\left(\frac{\pi_1 + \pi_2}{\pi_3}\right). \end{aligned}$$

2.1.3 Modelos Log-lineares

Há situações onde a variável resposta é representada por contagens, que podem constituir entradas numa tabela de contingência. Em geral, a relação ou associação entre duas variáveis categóricas é feita através da contribuição de cada célula nos totais marginais da tabela. A existência da associação entre as variáveis é, então, identificada por meio de um teste de independência, como o de qui-quadrado.

Os modelos log-lineares constituem uma classe de modelos lineares generalizados que utiliza o logaritmo como função de ligação. Esses modelos são, geralmente, usados para explicar a distribuição de freqüências numa tabela de contingência, principalmente quando há pelo menos duas variáveis dependentes, analisando a associação e a interação entre essas variáveis (Agresti, 2002; Agresti, 2007).

Os modelos log-lineares utilizam as freqüências esperadas $\mu_{ij} = n\pi_{ij}$, ao invés da probabilidade π_{ij}. Para uma tabela de dupla entrada, e sob independência, essa distribuição possui valores esperados $\mu_{ij} = n\pi_{i\cdot}\pi_{\cdot j}$, $\forall i, j$. Aplicando o logaritmo, obtém-se o modelo log-linear de independência

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y,$$

onde o parâmetro λ_i^X representa o efeito na linha i. Isso significa que, quanto maior o valor desse parâmetro, maior será a freqüência esperada na linha i. Raciocínio análogo pode ser aplicado ao parâmetro λ_j^Y, que representa o efeito na coluna j (Agresti, 2007).

Caso haja dependência estatística entre as variáveis, deve-se avaliar o efeito da interação entre as variáveis. O modelo completo (ou saturado), para uma tabela de

contingência 2×2, é dado por

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \tag{2.3}$$

onde λ_{ij}^{XY} representa o efeito da interação entre as variáveis X e Y.

As tabelas de contingência com dimensões $I \times J$ apresentam apenas $(I-1)(J-1)$ parâmetros não redundantes. Atribuindo-se o valor zero para a última linha e para a última coluna, obtém-se o produto cruzado de $(I-1)(J-1)$ variáveis. Neste caso, os testes de independência, com $(I-1)(J-1)$ graus de liberdade, permitem avaliar se esses parâmetros são iguais a zero (Agresti, 2007).

O modelo 2.3 possui um total de IJ parâmetros, sendo composto por um único parâmetro constante λ; e o número de parâmetros não redundantes para os termos λ_i^X, λ_j^Y e λ_{ij}^{XY} são, respectivamente: $(I-1)$, $(J-1)$ e $(I-1)(J-1)$.

Os modelos saturados fornecem um ajuste perfeito para os dados por sua capacidade de descrever qualquer conjunto de frequências esperadas. Entretanto, os modelos não saturados são preferíveis por fornecerem ajustes mais suaves e interpretação mais simples (Agresti, 2007).

A seção 2.1.3.3 apresenta os modelos log-lineares gráficos como uma ferramenta útil na busca de modelos mais parcimoniosos permitindo ajustes mais adequados aos dados.

Os modelos log-lineares não fazem distinção entre a variável desfecho e as variáveis preditoras. Algumas relações importantes podem ser obtidas a partir desses modelos.

2.1.3.1 Relação entre o Modelo Log-linear e a Razão de Chances

A razão de chances (*odds ratio*) mede a associação entre variáveis numa tabela de contingência de dupla entrada. Essa medida é obtida partindo-se da chance de sucesso (*odds of success*), definida como $odds = \pi/(1-\pi)$, que assume valores não negativos, sendo π a probabilidade de sucesso. Um valor maior que 1 indica que a ocorrência de um sucesso é mais provável do que a de uma falha e representa o número esperado de sucessos para cada falha observada. A probabilidade de sucesso pode ser, assim, obtida em função da chance de sucesso: $p = odds/(odds + 1)$ (Agresti, 2007).

Para uma tabela de contingência 2×2, a razão de chances (θ) é expressa como o quociente entre as chances de sucesso nas linhas 1 e 2, i.é,

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

A razão de chances não se altera quando se transpõem os valores da tabela de contingência, transformando-se linhas em colunas e colunas em linhas. O mesmo

ocorre quando se consideram as colunas como variáveis resposta e as linhas como variáveis explicativas e vice-versa. Assim, é dispensável a identificação de uma variável resposta para estimação de θ. Portanto, quando ambas as variáveis forem consideradas desfecho, pode-se utilizar o conceito de probabilidades conjuntas para se obter o valor da razão de chances (Agresti, 2007):

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

A razão de chances pode ser obtida a partir dos parâmetros de associação λ_{ij}^{XY}. Seja um modelo para tabelas de contingência 2×2, cujo logaritmo da razão de chances é expresso por

$$\begin{aligned}
\log\theta &= \log\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} \\
&= \log\mu_{11} + \log\mu_{22} - \log\mu_{12} - \log\mu_{21} \\
&= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\
&\quad - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\
&= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}. \qquad (2.4)
\end{aligned}$$

Fazendo-se a exponenciação da equação 2.4, obtém-se razão de chances (Agresti, 2007):

$$\theta = \exp(\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}).$$

Observa-se, portanto, que a razão de chances é obtida a partir dos termos de interação λ_{ij}^{XY}. Pode-se observar ainda que, se $\lambda_{ij}^{XY} = 0$, o resultado da equação 2.4 também será igual a zero, indicando que as variáveis X e Y são independentes.

2.1.3.2 Relação entre o Modelo Log-linear e o Modelo Logístico

Os modelos Log-lineares buscam a relação entre variáveis independentes categóricas, enquanto o modelo Logístico descreve a associação entre a variável resposta e um conjunto de variáveis explicativas. O modelo Logístico constitui um caso especial do modelo Log-linear com resposta binária e pode ser obtido a partir da construção de logitos para a variável dependente do modelo Log-linear (Agresti, 2007).

Seja o modelo de associação para tabelas de contingência tridimensionais, onde Y representa uma variável resposta binária; e X e Z são variáveis independentes categóricas (Agresti, 2007):

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

Calculando-se o logito de $P(Y = 1)$ quando as variáveis explicativas X e Z são, respectivamente, dos níveis i e k:

$$\begin{aligned} logit\Big(P(Y=1)\Big) &= \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \log\left(\frac{P(Y=1|X=i,Z=k)}{P(Y=0|X=i,Z=k)}\right) \\ &= \log\left(\frac{\mu_{i1k}}{\mu_{i0k}}\right) = \log(\mu_{i1k}) - \log(\mu_{i0k}) \\ &= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}) \\ &\quad - (\lambda + \lambda_i^X + \lambda_0^Y + \lambda_k^Z + \lambda_{i0}^{XY} + \lambda_{ik}^{XZ} + \lambda_{0k}^{YZ}) \\ &= (\lambda_1^Y - \lambda_0^Y) + (\lambda_{i1}^{XY} - \lambda_{i0}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{0k}^{YZ}). \end{aligned} \tag{2.5}$$

Com relação aos termos entre parênteses da equação 2.5, observa-se que os que não dependem da variável resposta, Y, são cancelados. Restam portanto: o primeiro e o segundo termos, que não dependem das variáveis explicativas, X ou Z; o terceiro e o quarto, que dependem de X; e o quinto e o sexto, que dependem de Z. Desta forma, o logito de $P(Y = 1)$ assume a forma aditiva (Agresti, 2007)

$$logit\Big(P(Y=1)\Big) = \alpha + \beta_i^X + \beta_k^Z. \tag{2.6}$$

Essa aditividade representa a ausência de interação entre as variáveis categóricas. Os modelos log-lineares não fazem distinção entre variáveis explicativas e variável resposta, o que não ocorre no caso do modelo Logístico.

A Tabela 2.3 (Agresti, 2007) apresenta a equivalência entre os modelos log-lineares e os modelos logísticos para uma tabela de contingência de três dimensões, sendo todas as variáveis categóricas e Y, uma variável resposta binária para o modelo logístico.

Na notação de Agresti (2007), o "símbolo log-linear" reflete o termo de maior interação do modelo log-linear. E o "símbolo logístico" representa as variáveis explicativas, que exercem influência sobre a variável resposta.

Vale observar que, conforme mostrado na Equação 2.5, os termos que não dependem da variável resposta são eliminados do modelo logístico, uma vez que esse modelo não descreve a associação entre variáveis explicativas apenas, e sim, entre o conjunto dessas variáveis e o desfecho. Por este motivo, embora presentes no modelo log-linear, esses termos não aparecem no modelo logístico.

Ainda sobre a Tabela 2.3, observa-se que o termo de interação entre as variáveis independentes, (XZ), aparece em todos os modelos log-lineares. Por outro lado, tal interação não é apresentada nos modelos logísticos, que só consideram interações que envolvam a variável resposta (Y). O modelo log-linear (Y, XZ) indica que a variável Y independe tanto de X quanto de Z. O modelo log-linear (XY, YZ, XZ) corresponde ao modelo logístico apresentado na Equação 2.6. Já o modelo (XYZ) representa o modelo saturado.

Tabela 2.3: Equivalência entre os Modelos Log-linear e Logístico

Símbolo Log-linear	Modelo Log-linear	Modelo Logístico	Símbolo Logístico
(Y, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$	α	$(-)$
(XY, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	$\alpha + \beta_i^X$	(X)
(YZ, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$	$\alpha + \beta_k^Z$	(Z)
(XY, YZ, XZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$	$\alpha + \beta_i^X + \beta_k^Z$	$(X + Z)$
(XYZ)	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$	$\alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$	$(X * Z)$

2.1.3.3 Modelos Log-lineares Gráficos

O estudo dos modelos gráficos requer alguns conceitos importantes para a sua compreensão. Inicialmente, serão abordados os conceitos relacionados à teoria dos grafos.

Um grafo $G = (V, E)$ é uma estrutura composta por um conjunto finito de vértices e um conjunto finito de arestas. Cada par de vértices pode, ou não, estar ligado por uma aresta. Um grafo é dito orientado se houver uma ordenação entre os pares de vértices. Neste caso, cada par de vértices será ligado por uma seta, que indicará a ordenação.

Dois vértices X, $Y \in V$, conectados por uma aresta são considerados vértices adjacentes, representados por $X \sim Y$. Denomina-se caminho, de comprimento n, entre dois vértices X_0 e X_n, a uma sequência de vértices $X_0, X_1, \ldots, X_n$ tal que vale a relação de adjacência $X_{i-1} \sim X_i$, $i = 1, 2, \ldots, n$.

Um grafo é dito completo se houver uma aresta ligando cada par de seus vértices. Clique é um subgrafo completo, ou seja, é um subconjunto de vértices adjacentes entre si e totalmente conectados. Uma clique maximal é um subgrafo completo que não está contido em outra clique. Para maior compreensão desses conceitos, recomenda-se a referência Boaventura-Netto (2006).

Se A, B e C são três subconjuntos distintos de V, diz-se que C separa A e

B se todo o caminho do vértice A para o vértice B for interceptado pelo vértice C. Por exemplo, no grafo mostrado na figura 2.1, diz-se que $\{X\}$ separa $\{Y, W\}$ e $\{Z\}$. Da mesma forma, $\{W\}$ e $\{Z\}$ são separados por $\{X, Y\}$ (Wasserman, 2004). Os modelos gráficos utilizam os grafos para representar a relação de independência entre variáveis.

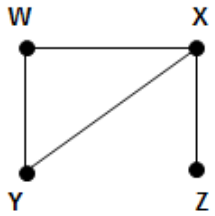


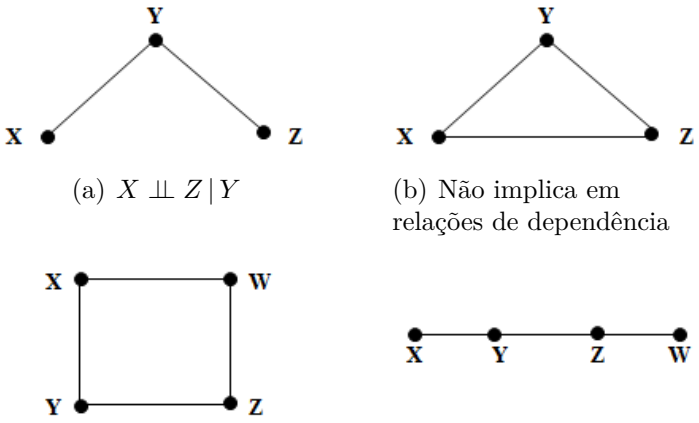
Figura 2.1: Relações de independência condicional entre variáveis

Geralmente, na modelagem de dados, é necessário que sejam fixadas algumas variáveis, que passam a representar uma parte constante do experimento, e analisar as variáveis restantes como se fossem variáveis aleatórias extraídas de uma densidade condicional, dadas as variáveis fixas (Edwards, 2000). Essas relações de independência condicional entre as variáveis podem ser visualmente identificadas através da modelagem gráfica, com o uso da notação utilizada por Dawid (1979).

Por definição, diz-se que duas variáveis aleatórias X, Y são condicionalmente independentes, dada uma terceira variável Z (simbolicamente, $X \perp\!\!\!\perp Y \,|\, Z$), se $f_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z), \forall x, y, z$. Ou seja, uma vez que se tenha conhecimento sobre Z, a variável Y não fornece nenhuma informação adicional sobre X. Pode-se, portanto, afirmar que $f_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z)$. Assim, são válidas as seguintes implicações (Wasserman, 2004):

$$
\begin{aligned}
X \perp\!\!\!\perp Y \,|\, Z &\implies Y \perp\!\!\!\perp X \,|\, Z \\
X \perp\!\!\!\perp Y \,|\, Z \quad \text{e} \quad U = h(X) &\implies U \perp\!\!\!\perp Y \,|\, Z \\
X \perp\!\!\!\perp Y \,|\, Z \quad \text{e} \quad U = h(X) &\implies X \perp\!\!\!\perp Y \,|\, (Z, U) \\
X \perp\!\!\!\perp Y \,|\, Z \quad \text{e} \quad X \perp\!\!\!\perp W \,|\, (Y, Z) &\implies X \perp\!\!\!\perp (W, Y) \,|\, Z \\
X \perp\!\!\!\perp Y \,|\, Z \quad \text{e} \quad X \perp\!\!\!\perp Z \,|\, Y &\implies X \perp\!\!\!\perp (Y, Z)\,.
\end{aligned}
$$

Define-se um grafo de Markov 2 a 2 como um grafo não orientado constituído de um conjunto de vértices, que representam as variáveis aleatórias, onde são omitidas as arestas entre pares de variáveis independentes, dado o restante das variáveis. Em outras palavras, segundo a propriedade de Markov 2 a 2, duas variáveis não adjacentes são condicionalmente independentes, dado o restante das variáveis. A Figura 2.2 ilustra esta definição, onde as relações de independência condicional 2 a 2 encontram-se representadas (Wasserman, 2004; Edwards, 2000).



(a) $X \perp\!\!\!\perp Z \,|\, Y$

(b) Não implica em relações de dependência

(c) $X \perp\!\!\!\perp Z \,|\, Y, W$ e $Y \perp\!\!\!\perp W \,|\, X, Z$

(d) A independência 2 a 2 implica em $X \perp\!\!\!\perp Z \,|\, Y, W$

Figura 2.2: Propriedade de Markov 2 a 2

Seja $G = (V, E)$ um grafo Markov 2 a 2. Sejam também A, B, e C subconjuntos distintos de V tais que C separa A e B. Nestas condições, tem-se $A \perp\!\!\!\perp B \,|\, C$, que representa as propriedades globais de Markov (Wasserman, 2004).

Uma forma mais simples de identificar as relações de independência entre as variáveis é obtida através do uso dos modelos log-lineares gráficos.

O uso dos grafos, em conjunto com os modelos log-lineares, é devido a Darroch et al. (1980), que desenvolveram uma conexão entre a teoria de Markov e os modelos log-lineares para definir e investigar os modelos gráficos para tabelas de contingência.

Segundo Darroch et al. (1980), os modelos gráficos podem ser representados por um grafo não-orientado finito, com tantos vértices quantas forem as dimensões da tabela de contingência, podendo também ser interpretados em termos de independência condicional.

Por definição, se (X_a, X_b, X_c) for um vetor aleatório, então $X_b \perp\!\!\!\perp X_c \,|\, X_a$ se e somente se todos os termos em λ (os λ-termos) da expansão log-linear, com uma ou mais coordenadas em b e em c, forem fixados em zero. Ou seja, um modelo é gráfico se todos os termos da expansão log-linear são diferentes de zero, exceto para os pares de coordenadas que não pertençam ao conjunto de arestas do grafo G. Assim, $\lambda_A(x) = 0$ se e somente se $\{i, j\} \subset A$ e (i, j) não forma uma aresta, caracterizando a relação de independência. Desta forma, muitas hipóteses podem ser geradas fixando-se em zero alguns λ-termos. O modelo gráfico será, então, determinado com o uso dos grafos e das independências condicionais 2 a 2 (Whittaker, 1990; Wasserman, 2004). Em outras palavras, para que um modelo log-linear seja gráfico é necessário que todas as cliques identificadas no grafo reflitam todas as interações presentes na expansão log-linear, bem como os termos correpondentes de ordem inferior.

Seja, por exemplo, $G = (5, 5)$ o grafo de 5 vértices formado pelo conjunto de arestas $E = \{(1,2),(2,3),(2,5),(3,4),(3,5),(4,5)\}$, apresentado na Figura 2.3. Trata-se de um modelo gráfico, pois os subscritos correspondentes às arestas omitidas no grafo não estão presentes na expansão log-linear (Whittaker, 1990; Wasserman, 2004).

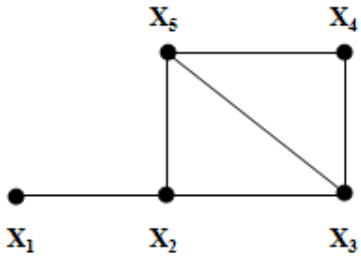


Figura 2.3: Grafo $G = (5, 5)$

O modelo gráfico correspondente, para $X = (X_1, X_2, X_3, X_4, X_5)$, possui a expansão log-linear dada pela equação 2.7.

$$\begin{aligned}\log f(x) = {} & \lambda_0(x) + \lambda_1(x) + \lambda_2(x) + \lambda_3(x) + \lambda_4(x) + \lambda_5(x) + \lambda_{12}(x) + \lambda_{23}(x) \\ & + \lambda_{25}(x) + \lambda_{34}(x) + \lambda_{35}(x) + \lambda_{45}(x) + \lambda_{235}(x) + \lambda_{345}(x). \end{aligned} \tag{2.7}$$

Seja agora a expansão log-linear dada pela equação 2.8. Apesar do grafo correspondente ao modelo ser o mesmo da Figura 2.3, o modelo é não gráfico, pois além das restrições de independência condicional, o modelo impõe restrições adicionais, que fixam em zero os termos de interação $\lambda_{235}(x)$ e $\lambda_{345}(x)$.

$$\begin{aligned}\log f(x) = {} & \lambda_0(x) + \lambda_1(x) + \lambda_2(x) + \lambda_3(x) + \lambda_4(x) + \lambda_5(x) + \lambda_{12}(x) + \lambda_{23}(x) \\ & + \lambda_{25}(x) + \lambda_{34}(x) + \lambda_{35}(x) + \lambda_{45}(x). \end{aligned} \tag{2.8}$$

Os modelos que apresentam outras restrições, além das de independência condicional, são denominados modelos hierárquicos e constituem uma classe mais ampla de modelos, onde é permitida maior flexibilidade pela atribuição do zero a alguns termos da expansão log-linear.

Um modelo log-linear é considerado hierárquico sempre que, se um λ-termo é fixado em zero, todos os termos de ordem mais elevada, e que possuam o mesmo conjunto de subscritos, sejam também restritos a zero. Ou seja, se $\lambda_a = 0 \implies \lambda_t = 0; \quad \forall a \subseteq t$ (Whittaker, 1990; Wasserman, 2004). Os modelos hierárquicos indicam a ausência de termos de independência condicionais, cuja presença é desnecessária no modelo. Um modelo log-linear hierárquico é gráfico se e somente se seus λ-termos maximais corresponderem às cliques no grafo de independência.

Seja o modelo $\log f(x) = \lambda_0(x) + \lambda_3(x) + \lambda_{12}(x)$. Este modelo possui o termo maximal $\lambda_{12}(x)$ e o seu grafo é dado pela Figura 2.4. Observa-se que $\lambda_1(x) = 0$, mas $\lambda_{12}(x) \neq 0$. Portanto, o modelo é não hierárquico.

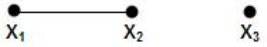
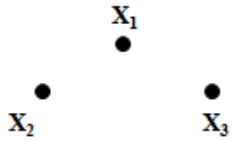


Figura 2.4: Grafo $G = (3, 1)$

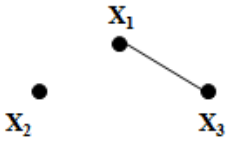
Os modelos log-lineares hierárquicos podem ser expressos através de geradores, que representam os termos de interação mais alta na expansão log-linear.

Por exemplo, o gerador $M = 1.2 + 1.3$ corresponde ao modelo $\lambda_0(x) + \lambda_1(x) + \lambda_2(x) + \lambda_3(x) + \lambda_{12}(x) + \lambda_{13}(x)$.

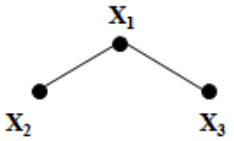
Vale observar que todo modelo gráfico é hierárquico, mas a recíproca nem sempre é verdadeira (Whittaker, 1990; Wasserman, 2004), como mostra a Figura 2.5 (Edwards, 2000). Dos dez modelos apresentados, todos são hierárquicos gráficos, com exceção de dois, que são não-gráficos: 2.5(d), pela ausência do termo de interação $\lambda_{123}(x)$; e 2.5(h), que não apresenta o termo $\lambda_{124}(x)$. Os modelos referentes às figuras 2.5(a) e 2.5(e) correspondem, respectivamente, ao modelo de interação mútua e ao modelo saturado.



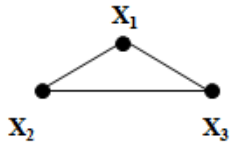
(a) $M = 1 + 2 + 3$



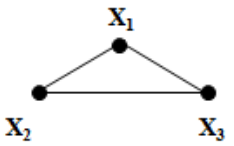
(b) $M = 2 + 1.3$



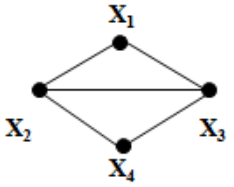
(c) $M = 1.2 + 1.3$



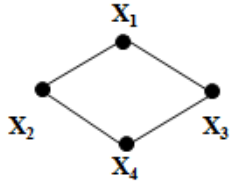
(d) $M = 1.2 + 1.3 + 2.3$



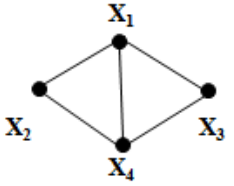
(e) $M = 1.2.3$



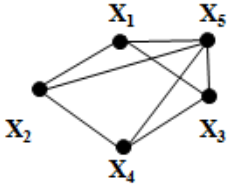
(f) $M = 1.2.3 + 2.3.4$



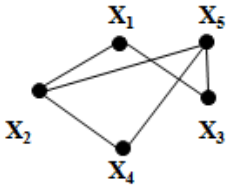
(g) $M = 1.2 + 1.3 + 3.4 + 2.4$



(h) $M = 1.2+2+4+ 1.3.4$



(i) $M = 1.2.5 + 1.3.5 + 2.4.5 + 3.4.5$



(j) $M = 1.2 + 1.3 + 2.4.5 + 3.5$

Figura 2.5: Modelos log-lineares hierárquicos

2.1.4 Modelo de Regressão de Poisson

Em estudos epidemiológicos, a variável resposta representa, em geral, o número de ocorrências de algum evento raro. Evento esse que pode ser, por exemplo, o número de novos casos de câncer de pulmão, observados em uma população, em deteminado período de tempo, sendo de interesse buscar a relação entre essas contagens e as variáveis independentes. A distribuição de Poisson é adequada para descrever tal fenômeno (Sahai & Khurshid, 1996), sendo definida por

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = exp(-\mu)\left(\frac{1}{y!}\right)exp(y\ \log\mu),$$

onde Y representa a variável dependente (contagens); μ é o seu valor esperado ($\mu = E(Y)$). Essa função de densidade tem a forma da família exponencial, que, na notação de Agresti (2002), é expressa por $f(y_i;\theta_i) = a(\theta_i)b(y_i)exp\Big[y_iQ(\theta_i)\Big]$, onde $\theta = \mu$; $a(\mu) = exp(-\mu)$; $b(y) = 1/y!$; e $Q(\mu) = \log(\mu)$.

Seja Y uma variável resposta, que segue uma distribuição de Poisson. Essa variável, tem portanto, função de probabilidade, esperança e variância definidas, respectivamente, por

$$\begin{aligned} &f(y_j) = \frac{e^{-\mu_j}\mu_j^{y_j}}{y_j!}, \qquad y_j = 0,1,2,\ldots, \qquad \mu_j > 0 \\ &E(Y_j) = \mu_j, \\ &Var(Y_j) = \mu_j. \end{aligned} \tag{2.9}$$

No modelo de regressão de Poisson, modela-se o logaritmo neperiano da média da variável dependente, obtendo-se assim,

$$\begin{aligned} \log(\mu_j) &= \beta_0 + \beta_1 x_{1j} + \ldots + \beta_k x_{kj}, \\ \mu_j &= exp(\beta_0 + \beta_1 x_{1j} + \ldots + \beta_k x_{kj}), \\ \mu_j &= \mu_0\ exp(\beta_1 x_{1j} + \ldots + \beta_k x_{kj}). \end{aligned} \tag{2.10}$$

A função de ligação para esse modelo pode ser expressa por

$$g(\mu_j) = x_i'\beta = \log(\mu_j), \qquad \text{onde} \quad x_j'\beta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

Essa função de ligação garante a obtenção de valores não negativos na predição de y_i. O método da máxima verossimilhança pode ser utilizado para a estimação dos parâmetros β's (Montgomery, 2001):

$$L(b_0, b_1, \ldots, b_n) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \frac{e^{-\sum_{i=1}^{n}\mu_i}\prod_{i=1}^{n}\mu_i^{y_i}}{\prod_{i=1}^{n} y_i!}$$

$$ln\Big(L(b_0, b_1, \ldots, b_n)\Big) = ln\Big(e^{-\sum_{i=1}^{n}\mu_i}\Big) + ln\Big(\prod_{i=1}^{n}\mu_i^{y_i}\Big) - ln\Big(\prod_{i=1}^{n} y_i!\Big) = -\sum_{i=1}^{n}\mu_i + \sum_{i=1}^{n} y_i\, ln(\mu_i) - \sum_{i=1}^{n} ln(y_i!).$$

2.2 Análise de Sobrevida

A técnica da análise de sobrevida é o método mais frequentemente utilizado para estudo de dados de naturezas diversas, desde a medicina, epidemiologia e saúde ambiental, até a criminologia, *marketing* e astronomia (Lee & Go, 1997).

A análise de sobrevida busca estudar fenômenos, onde a variável de interesse (desfecho) é o tempo até que uma falha no sistema seja observada. A presença de censuras é comum neste tipo de estudo, caracterizando-se pela impossibilidade de observar a ocorrência da falha. Neste caso, sabe-se apenas que o tempo de sobrevida é superior ao tempo em que foi observada a censura.

Na pesquisa clínica, por exemplo, a falha pode ser a morte ou um evento não fatal ou a recidiva de uma doença. A censura, por sua vez, pode ser representada pela falta de acompanhamento do paciente, em virtude de desistência do tratamento ou mudança de cidade, ou ainda, a morte por motivo alheio ao objeto do estudo. Serão, também, considerados censurados todos os dados relativos aos indivíduos que permanecerem no estudo após o período de investigação.

É comum na pesquisa clínica, a investigação do tempo de vida relacionado a uma única causa de morte. Entretanto, há modelos mais complexos em que a morte do indivíduo esteja relacionada a uma, dentre várias possíveis causas apontadas no estudo. Esses modelos são denominados modelos de riscos competitivos. Como referência importante, recomenda-se Kalbfleisch & Prentice (1980).

Basicamente, três funções são utilizadas em dados de sobrevivência: a função de sobrevida, a função de distribuição acumulada e a função de risco.

A função de sobrevida $S(t)$ representa a probabilidade de um indivíduo sobreviver além de um tempo t:

$$S(t) = P(T > t),$$

onde T é uma variável aleatória não-negativa, que mede o tempo até a ocorrência de falha ou de censura.

O complementar da função de sobrevida é a função de risco acumulada, que indica a probabilidade de uma falha ser observada até o tempo t:

$$F(t) = P(T \leq t) = 1 - S(t).$$

A função de risco (Cox & Oakes, 1984) representa o risco instantâneo da ocorrência de falha num pequeno intervalo de tempo Δt, sendo definida por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)},$$

onde $f(t)$ representa a função de densidade de probabilidade.

A modelagem de dados de sobrevida com observações censuradas deve incluir uma variável indicadora de censura, que assume o valor 1, caso a falha tenha sido observada, ou 0, em caso contrário (censura).

A presença de covariáveis é comum em dados de sobrevivência. Nesse caso, é interessante analisar o efeito das covariáveis sobre o tempo de sobrevida, bem como investigar se a interação entre elas é significante. O Modelo de Riscos Proporcionais de Cox (1972) pode ser utilizado para investigar tais associações. Esse modelo assume que, para um indivíduo com um vetor z de covariáveis, a função de risco num dado tempo t é dada por

$$h(t|z) = e^{\alpha' z} h_0(t),$$

onde $h_0(t)$ é a função de risco basal no tempo t para $z = 0$, ou seja, para todas as covariáveis $h(t) = h_0(t)$, quando $z = 0$; e α é um vetor de coeficientes desconhecidos relativo ao efeito das covariáveis.

O modelo de riscos proporcionais (RP) assume que o vetor z tenha um efeito multiplicativo na função de risco, impondo proporcionalidade entre as funções de risco e, consequentemente, uma limitação por não permitir o cruzamento das curvas. Como alternativa a esse modelo, surgiu o modelo de tempo de falha acelerado (FA) (Kalbfleisch & Prentice, 1980):

$$h(t|z) = e^{\alpha' z} h_0(t\, e^{\alpha' z}).$$

Posteriormente, um modelo híbrido (RP/FA) foi apresentado por Ciampi & Etezadi-Amoli (1985 apud Louzada-Neto & Pereira, 2000):

$$h(t|z) = e^{\alpha_1' z} h_0(t\, e^{\alpha_2' z}),$$

onde α_1 e α_2 são vetores de coeficientes a serem estimados.

Um modelo geral, considerando heterocedasticidade para indivíduos sob diferentes níveis de covariáveis, foi proposto por Louzada-Neto (1997). Esse modelo é denominado Modelo de Risco Híbrido Estendido (ERP/EFA) e permite que o parâmetro de forma seja dependente das covariáveis por meio de uma função heterocedástica (apud Louzada-Neto & Pereira, 2000).

$$h(t|z) = e^{\alpha_1' z} e^{\gamma' z} exp(\alpha_1' z)^{exp(\gamma_1' z)-1} h_0(t\, e^{\alpha_2' z})^{exp(\gamma' z)},$$

onde α_1, α_2 e γ são vetores de coeficientes de regressão desconhecidos e $h_0(.)$ é a função de risco basal. A relação de heterocedasticidade entre as covariáveis e o parâmetro de forma é dada pela função $exp(\gamma' z)$.

Casos particulares para o modelo ERP/EFA são obtidos quando: $\gamma = 0$ (modelo híbrido RP/FA); ou $\gamma = 0$ e $\alpha_2 = 0$ (modelo FA).

O modelo de risco estendido (ES) constitui uma versão mais geral do modelo ERP/EFA, com a introdução de um vetor de coeficientes de regressão γ_2', tendo sido proposto por Louzada-Neto (1997 e 1999 apud Louzada-Neto & Pereira, 2000):

$$h(t|z) = e^{\alpha_1' z} e^{\gamma_1' z} e^{\gamma_2' z} exp(\alpha_1' z)^{exp(\gamma_1' z)-1} h_0(t\, e^{\alpha_2' z})^{exp(\gamma_2' z)}.$$

Louzada-Neto & Pereira (2000) abordam algumas dificuldades envolvidas na modelagem de dados de sobrevida na presença de censuras, variáveis explicativas e riscos competitivos. Os autores formulam diversos modelos baseados na função de risco e estabelecem comparações entre eles, destacando a sua evolução desde o modelo de riscos proporcionais (RP) de Cox (1972) até o modelo de risco estendido (ES). Essa evolução é apresentada na Figura 2.6. Os modelos ERP e EFA constituem versões estendidas dos modelos RP e FA, respectivamente, para os casos particulares onde $\gamma \neq 0$ e $\alpha_2 = 0$ (modelo ERP) e $\alpha_1 = \alpha_2$ (modelo EFA).

A análise de sobrevivência parte, geralmente, do pressuposto que os tempos de sobrevida dos indivíduos sejam independentes. Entretanto, há casos em que essa suposição pode não ser válida, como por exemplo, no estudo envolvendo grupos ou conglomerados de indivíduos, onde os tempos de sobrevida dentro de cada grupo são mutuamente independentes. Estudos relativos a indivíduos sujeitos a eventos recorrentes, tais como ataques cardíacos ou ataques epilépticos ou, ainda, múltiplas seqüelas em pacientes com doenças crônicas, constituem casos onde múltiplos tempos de sobrevida são observados para cada indivíduo, caracterizando um modelo de sobrevivência multivariado. Nestas situações, é razoável supor que haja dependência entre os tempos de sobrevida de cada indivíduo (Colosimo & Giolo, 2006).

O modelo, geralmente, utilizado para descrever uma possível associação entre múltiplos tempos de sobrevivência para cada indivíduo é o denominado Modelo de

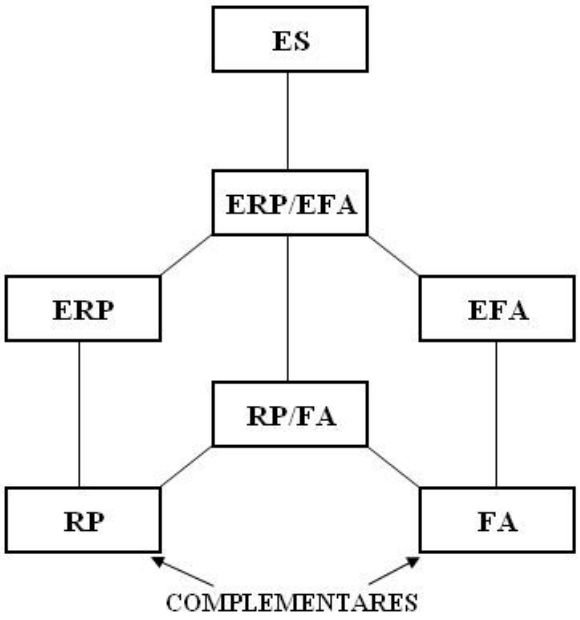


Figura 2.6: Evolução desde o modelo de Riscos Proporcionais até o de Risco Estendido

Fragilidade (Vaupel et al., 1979). Nesse modelo, um efeito aleatório (fragilidade) é introduzido na função risco, como um fator multiplicativo. Assim, o modelo de fragilidade, para o i-ésimo indivíduo, é definido por (Colosimo & Giolo, 2006)

$$\lambda_i(t) = z_i \lambda_0(t) exp(x_i' \beta),$$

onde β é um vetor de parâmetros desconhecidos; x_i é um vetor de covariáveis; $\lambda_0(t)$ é a função de risco basal, não especificada; e $z_1, \ldots, z_n$ são as fragilidades, que consistem em uma variável aleatória de fragilidade Z_i, independentes e identicamente distribuídas (iid), seguindo uma distribuição de probabilidade conhecida, com média igual a 1 e variância desconhecida.

A distribuição gama tem sido muito utilizada para modelar as variáveis de fragilidade Z_j, sobretudo pela sua conveniência algébrica. A função de densidade de probabilidade de Z_j, variáveis aleatórias independentes, para $Z_j \sim \Gamma(\eta, \upsilon)$, $\eta \geq 0$, $\upsilon \geq 0$ e considerando $\eta = \upsilon = \xi^{-1}$ é definida por (Colosimo & Giolo, 2006)

$$f(z) = \frac{\left(\frac{1}{\xi}\right)^{1/\xi}}{\Gamma\left(\frac{1}{\xi}\right)} \; z^{\frac{1}{\xi}-1} \; exp\Big\{ - z/\xi \Big\},$$

onde $z \geq 0$. Nestas condições, tem-se que $E(Z_j) = 1$ e $Var(Z_j) = \xi$.

Alguns casos especiais do modelo de fragilidade devem ser observados: (1) se a variância ξ for nula, o modelo de fragilidade reduz-se ao modelo de Cox; (2) indivíduos ou grupos, cuja fragilidade seja maior que 1 estão mais propensos a experimentar o evento de interesse; e (3) indivíduos com fragilidade menor que 1

terão maiores tempos até a ocorrência do evento de interesse (Carvalho et al., 2005).

No caso univariado, a fragilidade mede a heterogeneidade entre os indivíduos. No caso multivariado, além de explicar essa heterogeneidade, o modelo de fragilidade permite analisar o efeito de covariáveis latentes, ou seja, covariáveis que não tenham sido observadas e, portanto, não incluídas no estudo (Colosimo & Giolo, 2006). Outras distribuições de probabilidade, apontadas pela literatura e utilizadas na modelagem de variáveis aleatórias de fragilidade, são apresentadas nas publicações de Carvalho et al. (2005) e Colosimo & Giolo (2006).

A incorporação do efeito aleatório no modelo tem, como conseqüência, a alteração da estimativa pontual, proporcionalmente ao seu grau de importância. Desta forma, os efeitos das variáveis explicativas observadas e os intervalos de confiança para os respectivos parâmetros tendem a aumentar na medida em que aumenta o efeito aleatório. O fato de produzir um melhor ajuste no modelo não implica, entretanto, numa maior compreensão quanto aos mecanismos que influenciam a sobrevida dos pacientes. Por outro lado, proporciona uma maior confiança nas estimativas obtidas (Carvalho et al., 2005).

Como foi mencionado anteriormente, o modelo de Cox é, geralmente, utilizado para avaliar o efeito das covariáveis no tempo de sobrevivência. Entretanto, a suposição de riscos proporcionais, desse modelo, implica em que todos os indivíduos estejam suscetíveis ao evento de interesse, como por exemplo, a recidiva de uma doença. Na prática, porém, alguns indivíduos não experimentarão tal evento, ou seja, esses indivíduos estarão "curados". Isso indica que há um componente de cura no modelo (Price & Manatunga, 2001).

Historicamente, segundo Price & Manatunga (2001), os modelos de Boag (1949) e Berkson & Gage (1952) têm sido utilizados para estimar a fração de cura.

Price & Manatunga (2001) propuseram o uso de modelos de fragilidade para modelagem de dados de sobrevida com a inclusão da fração de cura.

Os modelos que incluem a fração de cura são denominados Modelos de Longa Duração. Rodrigues et al. (2008) sugeriram a unificação da Análise de Sobrevivência e Modelos de Longa Duração, de modo a integrar a teoria freqüentista com a teoria bayesiana numa abordagem de riscos competitivos. Segundo os autores, essa unificação seria obtida por meio de uma composição da função geradora de probabilidades do número de causas de ocorrência do evento de interesse (Feller, 1968) e a função de sobrevivência dos pacientes expostos ao risco.

A inclusão de indivíduos "curados" no modelo, isto é, não expostos ao risco de experimentarem o evento de interesse, conduz a superestimação da probabilidade de sobrevida e, conseqüentemente, a inferência incorretas (Price & Manatunga, 2001).

As estimativas dos tempos de sobrevida podem ser feitas através de modelos paramétricos, nos quais se assume que os dados seguem uma distribuição de proba-

bilidade, ou por modelos não paramétricos. Neste último caso, merecem destaque os estimadores de Kaplan-Meier e o de Nelson-Aalen.

2.2.1 Estimador de Kaplan-Meier

O estimador limite-produto de Kaplan-Meier, ou simplesmente, Estimador de Kaplan-Meier (Kaplan & Meier, 1958) é o mais utilizado e oferece maior vantagem computacional por estar disponível em diversos pacotes estatísticos (Colosimo & Giolo, 2006). Este estimador faz uma estimativa do tempo de sobrevivência considerando que a probabilidade de um indivíduo sobreviver até um determinado tempo t independe da probabilidade de sobrevida até cada um dos tempos precedentes.

Utilizando a notação adotada por Colosimo & Giolo (2006), define-se:

i) a probabilidade de ocorrência de morte no intervalo de tempo entre t_j e t_{j-1}, $j = 1, 2, \ldots, k$, dado que o indivíduo sobreviveu além do tempo imediatamente anterior:

$$q_j = P\Big(T \in [t_{j-1}, t_j) | T \geq t_{j-1}\Big);$$

ii) a probabilidade de sobrevida até um determinado tempo t_j:

$$S(t_j) = (1 - q_1) \cdot (1 - q_2) \cdot \ldots \cdot (1 - q_j);$$

iii) e, finalmente, a expressão geral do estimador de Kaplan-Meier:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right),$$

onde n_j representa o número de indivíduos em risco e d_j é o número de falhas observadas no instante t_j.

2.2.2 Estimador de Nelson-Aalen

O estimador de Nelson-Aalen (Nelson, 2000; Aalen, 1978) é um estimador não paramétrico que pode ser usado para estimar a função acumulada de risco, além de permitir que seja verificado graficamente o ajuste de modelos paramétricos (Aalen et al., 2008).

A taxa de risco acumulado $\Lambda(t)$ pode ser estimada, sem que sejam necessárias suposições com relação à função de risco $\alpha(t)$. Para a estimação de $\Lambda(t)$, o intervalo $[0, t)$ é particionado em pequenos intervalos assumindo-se que o evento ocorra, no máximo, uma vez em cada partição. Considerando que $\Big[s, s + d(s)\Big)$ seja um desses subintervalos, a contribuição do risco acumulado $\Lambda(t)$ sobre $\Big[s, s + d(s)\Big)$ é $\alpha(s)ds$, sendo expressa por

$$\Lambda(t) = \int_0^t \alpha(s)ds. \tag{2.11}$$

A equação 2.11 representa a probabilidade condicional de ocorrência do evento no intervalo $\Big[s, s + d(s)\Big)$, dado que não tenha ocorrido antes do tempo s. A estimativa de $\alpha(s)ds$ será zero, caso nenhuma ocorrência seja observada em $\Big[s, s + d(s)\Big)$. Entretanto, se o evento ocorrer no tempo $T_j \in [s, s + ds)$, a estimativa de $\alpha(s)ds$ será

$$\frac{1}{Y(s)} = \frac{1}{Y(T_j)},$$

onde $Y(t)$ representa o número de pessoas em risco no tempo t.

Incorporando-se essas contribuições sobre todos os intervalos, obtém-se o estimador de Nelson-Aalen, expresso por 2.12, sendo uma função escada crescente, contínua à direita, com incrementos de $1/Y(T_j)$ nos tempos observados (Aalen et al., 2008).

$$\tilde{\Lambda}(t) = \sum_{T_j \leq t} \frac{1}{Y(T_j)}. \tag{2.12}$$

O estimador de Nelson-Aalen apresenta propriedades semelhantes ao de Kaplan-Meier. A função de sobrevivência obtida a partir do estimador de Nelson-Aalen é dada por (Colosimo & Giolo, 2006)

$$\tilde{S}(t) = exp\{-\tilde{\Lambda}(t)\}.$$

Bohoris (1994 apud Colosimo & Giolo, 2006) provou que $\tilde{S}(t) \geq \hat{S}(t)$, onde $\hat{S}(t)$ é o estimador de Kaplan-Meier. Essas estimativas para a função de sobrevivência são bem próximas para amostras suficientemente grandes. Entretanto, para amostras pequenas, o uso do estimador de Nelson-Aalen é preferível (Carvalho et al., 2005).

A função de taxa de falha acumulada $\Lambda(t)$ é útil na escolha de modelos, embora não apresente interpretação em termos probabilísticos (Colosimo & Giolo, 2006).

2.2.3 Comparação de Curvas de Sobrevida

Na pesquisa clínica, com dados de sobrevida, é importante avaliar se um tratamento experimental é capaz de reduzir a mortalidade ou aumentar a sobrevida dos pacientes. Essa avaliação pode ser feita através da comparação da curva de Kaplan-Meier, correspondente ao grupo tratado com a nova terapia, com a do outro grupo, onde um tratamento convencional ou um placebo tenha sido utilizado.

Vários testes estatísticos têm sido propostos para identificar se as diferenças observadas nas curvas de Kaplan-Meier são significantes. Dentre os diversos testes não-paramétricos existentes, encontram-se os testes Log-rank (Cox, 1972; Mantel, 1966; Peto & Peto, 1972), Mantel-Haenszel (Mantel & Haenszel, 1959), Gehan (Gehan,

1965; Breslow, 1970), Tarone-Ware (Tarone & Ware, 1977) e Peto-Prentice (Peto & Peto, 1972; Prentice, 1978). Testes esses, que diferem apenas pela ponderação adotada.

Cada teste tem como base uma tabela de contingência para cada tempo, t, onde a ocorrência do evento tenha sido observada. Essa tabela leva em consideração o número de mortes (falhas) e o número de indivíduos em risco em cada grupo.

O teste Log-rank, também conhecido como teste Mantel-Cox, é o método mais amplamente utilizado para a comparação de curvas de sobrevida (Machin et al., 2006). Sob a hipótese nula de igualdade das funções de sobrevida, o teste log-rank segue aproximadamente uma distribuição χ^2, com $(g-1)$ graus de liberdade, onde g é o número de grupos em estudo.

A expressão geral para o teste Log-rank (LR) é apresentada através da equação 2.13 (Machin et al., 2006), onde O_j é o número observado e E_j é o número esperado de mortes, no tempo j:

$$\chi^2_{LR} = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j}. \tag{2.13}$$

A tabela de contingência a ser utilizada, em cada instante t, para comparação de duas curvas de sobrevivência relativas a dois tipos de tratamento, A e B, é mostrada na Tabela 2.4 (Machin et al., 2006).

Tabela 2.4: Tabela de contingência para testes de igualdade de curvas de sobrevida

Tratamento	Mortos	Vivos	Total
A	a_t	c_t	m_t
B	b_t	d_t	n_t
Total	r_t	s_t	N_t

Na notação de Machin et al. (2006), o número esperado de mortes, no instante t, assumindo-se uma distribuição hipergeométrica, é dado por

$$E_{At} = \frac{r_t m_t}{N_t} \quad \text{e} \quad E_{Bt} = \frac{r_t n_t}{N_t}; \tag{2.14}$$

e a sua variância, calculada para cada um dos tempos t, é expressa por

$$V_t = \frac{m_t n_t r_t s_t}{N_t^2 (N_t - 1)}. \tag{2.15}$$

O teste de Mantel-Haenszel (MH) é uma versão do teste log-rank que leva em

consideração a variância, V_t. Assim como o teste LR, o teste MH segue, aproximadamente, uma distribuição χ^2, com 1 grau de liberdade, quando são comparadas duas curvas de sobrevida. O teste MH é definido como (Machin et al., 2006)

$$\chi^2_{MH} = \frac{(O_A - E_A)^2}{\sum V_t}, \tag{2.16}$$

onde $O_A = \sum O_{At}$; e $E_A = \sum E_{At}$.

O uso do teste log-rank é recomendado quando são satisfeitas as suposições de riscos proporcionais. Entretanto, caso tais suposições sejam violadas, poderá haver uma perda de poder do teste. Nesse caso, o uso dos chamados testes Mantel-Haenszel ponderados é aconselhável (Machin et al., 2006).

Os testes Mantel-Haenszel ponderados (MHw), para a comparação da função de sobrevida de dois grupos, A e B, utilizam basicamente a equação 2.17, assumindo-se uma distribuição χ^2, com 1 grau de liberdade

$$\chi^2_{MHw} = \frac{\sum w_t(O_{At} - E_{At})^2}{\sum w_t^2 V_t}, \tag{2.17}$$

onde E_{At} e V_t são definidas, respectivamente, pelas equações 2.14 e 2.15. Os pesos w_t podem variar com o tempo, sendo definidos conforme mostra a Tabela 2.5, onde R_t é o total de pacientes em risco no tempo t; e d_j é o número observado de mortes (Machin et al., 2006).

Tabela 2.5: Testes Não-paramétricos para comparação de curvas de sobrevida

Teste	Peso
Mantel-Haenszel	$w_t = 1$
Gehan	$w_t = R_t$
Tarone-Ware	$w_t = \sqrt{R_t}$
Peto-Prentice	$w_t = \hat{S}(t) = \prod_{j=1}^{t} \frac{R_j d_j + 1}{R_j + 1}$

Se $w_t = 1$, $\forall t$, todos os tempos onde ocorrerem as mortes serão ponderados igualmente, transformando-se no teste Mantel-Haenszel sem ponderação (equação 2.16).

Por outro lado, se $w_t = R_t$, tem-se o teste Gehan, também conhecido como teste Breslow ou Wilcoxon generalizado. Na maioria dos casos, R_t diminui com o passar do tempo, em virtude do aumento do número de mortes e dos casos censurados. Por este motivo, o teste Gehan atribui maior ponderação aos tempos anteriores das

curvas de sobrevida. Esse tipo de ponderação pode utilizado para avaliar se o efeito na sobrevida é mais acentuado nas fases iniciais de um tratamento, diminuindo com o passar do tempo (Machin et al., 2006; Kleinbaum & Klein, 2005).

O teste Tarone-Ware (TW) assume pesos iguais a $\sqrt{R_t}$, ponderando mais os eventos ocorridos nos tempos iniciais. O teste TW é menos restritivo que o teste Gehan (Machin et al., 2006).

O teste Peto-Prentice, que pondera pela função de sobrevida, dá mais ênfase aos tempos inciais, tendo como vantagem não ser influenciado por padrões de censura inerentes a cada grupo em estudo (Machin et al., 2006).

2.3 Árvores de Decisão

Os modelos não paramétricos de regressão, com múltiplas variáveis explicativas, sofrem com o problema da "praga da dimensionalidade". Ou seja, se o problema em análise envolver um número muito grande de variáveis independentes, torna-se difícil capturar as características relevantes, como por exemplo, a influência das interações entre variáveis explicativas. Modelos paramétricos, como os modelos lineares generalizados ou os modelos aditivos generalizados, buscam identificar os principais efeitos nas variáveis dependentes e descartar quaisquer interações locais ou globais dessas variáveis. Por outro lado, resultados obtidos por técnicas tais como *projection pursuit regression* e redes neurais podem ser de difícil interpretação. O método CART (*classification and regression trees*) pode ser útil nestes tipos de situação (Rao et al., 2008).

Breiman et al. (1984) introduziram o método CART, um algoritmo de particionamento recursivo do espaço de covariáveis, fornecendo como resultado um modelo estruturado em árvore. Caso a variável independente seja do tipo contínuo, o modelo é denominado árvore de regressão. Se a variável desfecho for binária, o modelo logístico, apresentado na Seção 2.1.1, pode ser implementado. Neste caso, onde a variável resposta é binária, o método CART implementa a árvore de classificação (Rao et al., 2008), próximo assunto a ser abordado.

2.3.1 Árvores de Classificação

O problema de classificação utiliza, em geral, um conjunto de medidas com a finalidade de predizer, de forma sistemática, a classe à qual o objeto ou caso em estudo pertence.

Seja o espaço de medidas χ, contendo todos os possíveis vetores de medidas, e $C = \{1, 2, \ldots, J\}$ as classes associadas aos objetos a serem classificados. A cada

vetor de medidas $x \in \chi$ é atribuída uma classe em C, de acordo com uma regra de atribuição.

Um classificador ou regra de classificação é definido através de uma função $d(x) = j$, que associa cada vetor x a um ponto no espaço χ. Desta forma, define-se

$$A_j = \{x; d(x) = j\},$$

onde $A_1, A_2, \ldots, A_j$ são disjuntos e $\chi = \bigcup_j A_j$, tal que j é a classe predita, $\forall x \in A_j$ (Breiman et al., 1984).

A construção de um classificador tem, como base, o conhecimento anterior, representado por uma amostra de aprendizagem $\mathcal{L}$, definida como

$$\mathcal{L} = \{(x_1, j_1), \ldots, (x_N, j_N)\},$$

onde $(x_1, j_1), \ldots, (x_N, j_N)$ são os dados observados.

O estudo de classificação pode ter como proposta básica a obtenção de um classificador acurado ou a identificação da estrutura preditiva do problema. Neste caso, busca-se entender quais variáveis, ou interações de variáveis, são responsáveis pelo fenômeno em análise.

Algumas técnicas de estatística vêm sendo desenvolvidas para solucionar os problemas de classificação. O método CART (Breiman et al., 1984) implementa o modelo de classificação com o uso de árvores de decisão.

Os classificadores estruturados em árvores binárias são construídos através de divisões sucessivas, de subconjuntos do espaço de medidas χ, em dois subconjuntos descendentes.

Para construção de uma árvore binária, são necessários três elementos fundamentais: uma amostra de aprendizagem $\mathcal{L}$; uma regra de partição dos nós; e um critério de parada de divisão. Cada subconjunto terminal t representa uma partição de χ. O classificador faz a predição da classe à qual o nó terminal pertence, atribuindo-o um rótulo (Breiman et al., 1984).

A idéia básica para a construção de uma árvore de classificação é fazer a partição de forma tal que os dados, em cada subconjunto de nós descendentes, sejam mais "puros" que seus antecessores. Assim, a implementação é realizada conforme os passos a seguir (Breiman et al., 1984):

1) Definir a probabilidade a priori, $p(j|t)$, $j = 1, 2, \ldots, N$, como a proporção dos casos $x_n \in t$, que tem origem na classe j e é tal que $\sum_{j=1}^{N} p(j|t) = 1$.

2) Definir uma medida $i(t)$, de impureza do nó t, como uma função ϕ, não negativa, de $p(j|t)$ de forma tal que ϕ seja máxima. A bondade da divisão é definida como a

redução da impureza no nó, sendo expressa por

$$\Delta_i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R), \tag{2.18}$$

onde s é uma divisão candidata do nó t, particionando-o em dois outros nós t_L (à esquerda) e t_R (à direita); p_L e p_R representam a proporção de casos em t associados a t_L e t_R, respectivamente, conforme mostra a Figura 2.7.

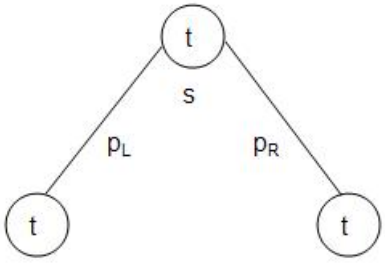


Figura 2.7: Construção da árvore de classificação

3) Definir um conjunto S de questões candidatas a divisões binárias s em cada nó, cuja resposta conduzirá a observação para o nó t_L ou para t_R.

O crescimento da árvore se dá, inicialmente, através de uma busca entre todos candidatos a divisões para identificar aquelas que fornecem o maior decréscimo na impureza, ou seja,

$$\Delta_i(s^*, t_1) = \max_{s \in S} \Delta_i(s, t_1),$$

onde s^* é o subconjunto de candidatos que produzem a maior redução da impureza do nó; e t_1 é o nó raiz.

O nó t_1 é, então, subdividido em outros dois nós t_2 e t_3. O mesmo procedimento de busca é aplicado em cada um dos nós resultantes e repetido, sucessivamente, nos nós subsequentes, até que uma regra heurística para interrupção do crescimento da árvore seja satisfeita.

Assim, o critério de parada de divisão será satisfeito quando nenhum decréscimo significante na impureza de um nó t for identificado. Nesse caso, o processo de repartição do nó será interrompido e o mesmo será considerado um nó terminal, ou seja uma folha da árvore.

As principais medidas de impureza são:
1) Taxa de erro de classificação, que por ser um indicador da qualidade não fornece nenhuma informação sobre a complexidade da árvore. Entretanto, esta medida possibilita avaliar a capacidade de predição do modelo (Tura, 2001). A taxa de erro de classificação no nó t é expressa por

$$R(t) = \frac{N_t}{N}\{1 - \max_j p(j|t)\},$$

onde N_t representa o número de observações no nó t e N é o número total de observações. A taxa de erro de classificação da árvore $R(T)$ é dada pelo somatório de R(t) em todos os nós terminais, assim,

$$R(T) = \sum_{t \in T} R(t).$$

2) Índice de Gini, usado para avaliar a probabilidade entre os nós, uma vez que reduz a importância das classes com probabilidade pequena em cada nó (Tura, 2001). O índice de Gini é dado por

$$i(t) = 1 - \sum_{j=1}^{J} p^2(j|t).$$

3) Entropia, que avalia as probabilidades das classes em cada nó, aumentando as diferenças entre as classes (Tura, 2001). A entropia é definida por

$$i(t) = -\sum_{j=1}^{J} p(j|t) \log p(j|t).$$

Um procedimento adequado para identificar o tamanho ótimo da árvore de classificação consiste em: (1) usar uma técnica de corte "*pruning*", ao invés de uma regra de parada de divisão, ou seja, fazer a árvore crescer o máximo possível e, então, podá-la; e (2) usar estimativas mais acuradas da probabilidade de erro de classificação, $R^*(t)$, para selecionar o tamanho correto da árvore dentre as subárvores podadas (Breiman et al., 1984).

O primeiro passo consiste, portanto, em fazer a árvore crescer até atingir o maior tamanho possível, permitindo que o processo de divisão continue até que todos os nós terminais sejam bem pequenos ou puros ou, ainda, contenham apenas vetores de medidas idênticas. Isso significa, continuar particionando até que cada nó terminal contenha uma única amostra (Breiman et al., 1984), o que seria inviável computacionalmente, para grandes amostras.

A característica da classe de um nó terminal é determinada pela regra da pluralidade (Breiman et al., 1984), ou seja, se

$$p(j_0|t) = \max_j P(j|t),$$

então t é declarado nó terminal da classe j_0.

Resumindo, partindo-se do nó raiz, a construção da árvore de tamanho máximo, pelo algoritmo CART, segue os seguintes passos (Yohannes & Webb, 1999):

1) Divide-se a primeira variável em todos os pontos possíveis de divisão. Para cada possível ponto de divisão, é feita uma pergunta do tipo "$x \leq \alpha$?", onde α é um limiar que define uma divisão específica do subconjunto X_t. Os casos, cuja resposta seja "sim", serão enviados para o nó à esquerda, caso contrário, para o nó à direita;

2) Para cada ponto de partição, aplica-se o critério de bondade da divisão e avalia-se a redução da impureza. A melhor divisão da variável é aquela que fornece a maior redução da impureza;

3) Repete-se os dois passos anteriores para cada uma das demais variáveis no nó raiz;

4) Ordena-se as melhores divisões, em cada variável, de acordo com a redução de impureza obtida, e seleciona-se a variável e o respectivo ponto de divisão, que proporciona a maior redução da impureza do nó raiz;

5) Atribui-se classes a esses nós, segundo a regra de minimização os erros de classificação;

6) Repete-se recursivamente os 5 passos anteriores, aplicando-os a cada nó não terminal;

7) O processo de divisão continua, construindo-se uma árvore de maior tamanho possível.

Entretanto, não é necessário partir de uma árvore de tamanho máximo, T_{max}, para que seja aplicada a técnica de corte. Basta que a árvore seja de tamanho suficientemente grande, no sentido que, se o processo tiver início a partir de $T_{max}^{'}$, será obtida uma subárvore contida em T_{max}. Desta forma, um processo iniciado com $T_{max}^{'}$ produzirá a mesma subárvore (Breiman et al., 1984).

Assim, pode-se especificar um número N_{min} e continuar o processo de particionamento até que cada nó terminal seja puro, ou $N_t \leq N_{min}$, ou contenha somente vetores de medidas idênticas. Em geral, N_{min} deve ser fixado em 5 e, eventualmente, em 1 (Breiman et al., 1984).

Portanto, independente do modo como T_{max} tenha sido construída ou do critério de divisão adotado, o processo de corte parte de uma árvore inicial $T_{max}^{'}$; calcula R_t para cada nó $t \in T_{max}^{'}$; e poda, progressivamente, $T_{max}^{'}$ no sentido do nó raiz, de forma tal que $R(t)$ seja o menor possível em cada estágio (Breiman et al., 1984).

2.3.2 Árvores de Regressão

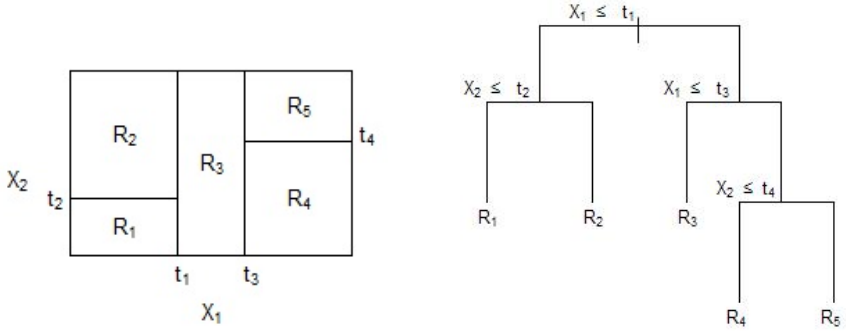
A análise de regressão envolve a construção de um preditor $d(x)$, definido sobre o espaço de medidas χ, a partir de uma amostra de aprendizagem $\mathcal{L}$.

Um preditor estruturado em árvore é semelhante ao classificador estruturado em árvore. Entretanto, não há necessidade de serem especificadas probabilidades a priori.

Nas árvores de regressão, o espaço de medidas χ é particionado por uma sequência de divisões binárias, onde, em cada nó terminal t, o valor predito da variável dependente $y(t)$ é constante (Breiman et al., 1984).

Por exemplo, seja um problema de regressão considerando o desfecho Y e as variáveis de entrada X_1 e X_2, todas definidas num intervalo unitário. Inicialmente, subdivide-se o espaço χ em duas subregiões e modela-se a resposta pela média de Y. A escolha da variável de entrada e do ponto de divisão é feita de modo a se obter o melhor ajuste. Uma ou ambas as subregiões são novamente subdivididas até que um critério de parada seja satisfeito (Hastie et al., 2009).

Na Figura 2.8(a), a variável X_1 foi, inicialmente, dividida no ponto $X_1 = t_1$. Em seguida, a subregião $X_1 \leq t_1$ foi dividida no ponto $X_2 = t_2$ e a região $X_1 > t_1$, por sua vez, no ponto $X_1 = t_3$. Por fim, dividiu-se a subregião $X_1 > t_3$ no ponto $X_2 = t_4$. Como resultado, obteve-se 5 subregiões $R_1, R_2, \ldots, R_5$. A Figura 2.8(b) apresenta a árvore binária referente a este modelo (Hastie et al., 2009).



(a) Partição do espaço em subregiões

(b) Árvore binária

Figura 2.8: Árvore de regressão

O modelo de regressão correspondente faz a predição utilizando a constante C_m na subregião R_m, obtendo-se, assim

$$\hat{y} = \sum_{m=1}^{5} C_m I\{(X_1, X_2) \in R_m\}.$$

Com mais de duas variáveis de entrada, é difícil visualizar a partição do espaço de caracteríscas na forma apresentada na Figura 2.8(a). Tal visualização torna-se bem mais fácil com o uso da árvore binária, tendo com principal vantagem, a facilidade de interpretação (Hastie et al., 2009).

O processo de construção de uma árvore de regressão é feita de forma semelhante ao da árvore de classificação. Partindo-se de uma amostra de aprendizagem $\mathcal{L}$, a determinação de um preditor em árvore envolve três elementos fundamentais: (1) a forma de selecionar uma divisão em cada nó intermediário; (2) uma regra para determinar um nó terminal; e (3) uma regra para atribuir o valor $y(t)$ a cada nó terminal t (Breiman et al., 1984).

O critério de divisão tem como base o menor valor obtido no cálculo da impureza do nó. Para a árvore de regressão a função de impureza, geralmente utilizada, é o erro quadrático médio, sendo expresso por

$$i(t) = \sum_{t} \frac{(y - \bar{y})^2}{n_t},$$

onde n_t é o número de elementos no nó t.

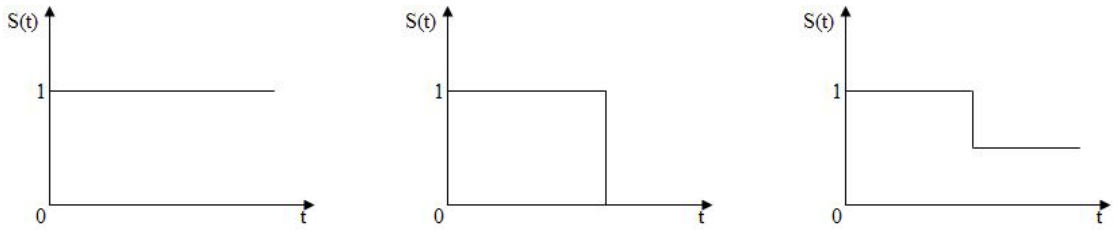
As principais etapas para a construção de uma árvore de regressão, com o uso do algoritmo CART, são os apresentados a seguir (Yohannes & Webb, 1999):
1) Partindo-se do nó raiz, faz-se todas as possíveis partições em cada uma das variáveis independentes;
2) Para cada divisão, calcula-se a medida de impureza do nó e o valor da redução da impureza;
3) Seleciona-se a melhor partição tendo, como base, o critério de bondade da divisão;
4) Para cada nó não terminal, repete-se, recursivamente, os três passos anteriores de modo a se obter a maior árvore possível;
5) Aplica-se a técnica do corte, progressivamente, no sentido do nó raiz e constrói-se uma sequência de subárvores, das quais uma árvore ótima será selecionada.

2.3.3 Árvore de Sobrevida

Os modelos baseados em árvore são amplamente utilizados em análise de sobrevivência, com dados censurados no tempo, sendo usados, particularmente, em aplicações biomédicas.

Em geral, duas abordagens são utilizadas para a construção de árvores de sobrevivência. A primeira usa uma medida de homogeneidade dentro do nó, por exemplo, a distância entre a estimativa das curvas de sobrevivência de Kaplan-Meier. A segunda abordagem é baseada em uma medida de separação entre os nós através de uma estatística de teste, para distinguir entre os tempos de sobrevivência (Hothorn et al. 2004).

Para medir a homogeneidade dentro do nó, pode-se utilizar o conceito de impureza do nó para dados censurados. Um nó pode ser considerado puro se todas as falhas ocorrerem simultaneamente. Neste caso, a curva de Kaplan-Meier pode assumir três possíveis aspectos (Zhang & Singer, 1999): todas as observações censuradas (Figura 2.9(a)); todas as falhas ocorrem ao mesmo tempo e nenhuma observação é censurada (Figura 2.9(b)); e todas as falhas ocorrem ao mesmo tempo seguidas por censuras (Figura 2.9(c)).



(a) (b) (c)

Figura 2.9: Curvas de Kaplan-Meier para um nó homogêneo

Gordon & Olshen (1985 apud Crowley et al., 1995) sugeriram o uso das métricas de Wasserstein para medir a variabilidade dentro de um nó. Essa variabilidade é definida como a menor distância entre o estimador produto limite para o nó e qualquer função escada. A métrica de Wasserstein calcula a distância horizontal entre funções de distribuição, sendo expressa como (Shorack & Wellner, 1986 apud Crowley et al., 1995)

$$\left[\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p du \right]^{1/p},$$

onde F_1 e F_2 são as funções de distribuição.

Assim, a impureza do nó t pode ser definida como (Zhang & Singer, 1999)

$$i(t) = \min_{\delta_S \in \mathcal{P}} d_p(S_t, \delta_S),$$

onde S_t é a curva de Kaplan-Meier dentro do nó t; δ_S representa melhor combinação entre S_t e uma das curvas pertencentes a $(\mathcal{P})$, apresentadas na Figura 2.9.

Um método alternativo consiste na utilização de uma estatística de teste para comparação de duas curvas de Kaplan-Meier, partindo do princípio que dois nós filhos relativamente puros, tendem a apresentar diferenças entre si. Assim, dois nós filhos diferentes podem ser obtidos aumentando-se a variação entre eles e, consequentemente, reduzindo-se a variação dentro desses nós, o que implica na homogeneidade dos nós (Zhang & Singer, 1999).

Ciampi et al. (1986) e Segal (1988) selecionaram a melhor divisão do nó através do maior valor obtido no teste log-rank (apud Zhang & Singer, 1999), método amplamente utilizado para testar a diferença entre os tempos de sobrevida, cujo cálculo foi apresentado na Seção 2.2.3. A bondade da divisão pode ser calculada através da equação 2.18.

2.4 Análise de Componentes Principais, Algoritmo EM e Critério de Akaike

Esta seção apresenta o resumo de alguns métodos estatísticos para análise de dados, estimação de parâmetros e escolha de modelos, que foram utilizados na análise de alguns dados clínicos, cujas aplicações encontram-se nos apêndices desta tese.

2.4.1 Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é um método multivariado, que parte de um conjunto de p variáveis originais, com o objetivo de: (1) obter um novo conjunto de variáveis não correlacionadas; (2) reduzir a dimensionalidade dos dados, de modo a obter um conjunto representativo e que retenha a maior quantidade possível da informação; e (3) a partir desse novo conjunto de variáveis, obter interpretação razoável para o fenômeno em estudo (Pereira & Rao, 2009).

É de interesse que haja redundâncias no conjunto de variáveis originais. Quanto maior a redundância, maior a correlação (positiva ou negativa) entre as variáveis e, consequentemente, melhores serão os resultados. Neste caso, o conjunto de variáveis originais poderá ser representado por um pequeno número de componentes principais (Manly, 2008).

A relação de dependência entre as variáveis é dada pela matriz de correlação ($\boldsymbol{R}$) ou pela matriz de covariâncias ($\boldsymbol{S}$). A técnica da Decomposição do Valor Singular (SVD) pode ser de grande utilidade na análise de componentes principais. A SVD pode ser aplicada sobre a matriz $\boldsymbol{R}$ ou $\boldsymbol{S}$, embora os resultados obtidos não sejam os mesmos. Como a PCA não é invariante a transformações, geralmente utiliza-se a matriz de correlação (Pereira & Rao, 2009). A SVD parte do princípio que toda a matriz de dados $\boldsymbol{X}$ pode ser decomposta na seguinte estrutura (Johnson & Wichern, 2007; Manly, 2008; Pereira & Rao, 2009):

$$\boldsymbol{X} = \boldsymbol{U}_{mm}\boldsymbol{\Lambda}_{mn}\boldsymbol{V'}_{nn},$$

onde $\boldsymbol{U}$ é uma matriz composta pelos autovetores ortonormais de $\boldsymbol{XX'}$; $\boldsymbol{V'}$ é uma matriz formada pelos autovetores ortonormais de $\boldsymbol{X'X}$; e $\boldsymbol{\Lambda}$ é uma matriz diagonal

cujos elementos são os autovalores de $\boldsymbol{U}$ ou de $\boldsymbol{V}$, dispostos em ordem crescente.

Seja $X = X_1, X_2, \ldots, X_p$ um conjunto de p variáveis originais. O i-ésimo componente principal, Z_i, é uma combinação linear das variáveis originais, sendo definido como (Manly, 2008)

$$Z_i = \sum_{j=1}^{p} a_{ij},$$

sujeito à restrição

$$\sum_{j=1}^{p} a_{ij}^2 = 1,$$

onde o coeficiente a_{ij} corresponde ao autovetor a_i, referente ao i-ésimo maior autovalor λ_i.

Para evitar redundâncias, cada Z_i deverá estar não correlacionado com os demais. Os componentes são, então, dispostos em ordem decrescente de variância, obtendo-se assim,

$$Var(Z_1) \geq Var(Z_2) \geq \ldots \geq Var(Z_p).$$

Como $\sum_{i=1} Var(Z_i) = \sum_{i=1} Var(X_i)$, os componentes principais, Z_i, concentram toda a variação existente nos dados originais (Manly, 2008). Assim, os componentes que explicarem apenas uma pequena parte da variação dos dados deverão ser descartados e os k primeiros componentes serão, então, considerados os componentes principais, por concentrarem grande parte dessa variação.

Não existe um modelo que determine o valor de k ideal para cada situação. Hair et al. (1998 apud Tura, 2001) apontam quatro possíveis alternativas para determinação desse valor: (1) determinação a priori, de acordo com o conhecimento prévio do especialista; (2) utilização do total da variância explicada; (3) seleção através dos autovalores, sendo geralmente consideradas as dimensões cujos autovalores sejam maiores ou iguais a 1; e (4) uso do *scree plot*, método gráfico onde o número de componentes são apresentados no eixo das abscissas e os autovalores são mostrados no eixo das ordenadas. O *scree plot* produz uma curva decrescente, que tende para uma reta horizontal. O valor da abscissa, correspondente ao ponto onde a essa reta horizontal se inicia, indica o número de componentes a serem utilizados.

2.4.2 Algoritmo EM

O Algoritmo EM (*Expectation-Maximization Algorithm*) é um método iterativo, proposto por Dempster et al. (1977), para a obtenção da estimativa de máxima verossimilhança quando há dados incompletos, ou quando for mais simples maximizar a verossimilhança em problemas onde quantidades desconhecidas estão envolvidas (Migon & Gamerman, 1999).

Sejam X um vetor de dados observados e Z um vetor de quantidades desconhecidas. Seja, também, $Y = (X, Z)$ o conjunto de dados completos, cuja função de densidade é dada por $p(y|\theta) = p(x, z|\theta)$, $\theta \in \Theta$.

Seja, ainda, $p(z|x, \theta)$ a densidade dos dados não observados, z, condicional a observação dos dados, x, que também depende de θ. Nessas condições, obtém-se a estimativa de máxima verossimilhança de θ, através da maximização da função de log verossimilhança marginal (Migon & Gamerman, 1999),

$$L(\theta; x) = \log\Big(\int p(x, z; \theta)dz\Big),$$

que envolve o cálculo da integral de grande dimensão, relacionada à marginalização de $p(x, z|\theta)$. Uma alternativa para esse cálculo consiste em considerar que a distribuição dos dados completos pode ser fatorada como $p(x, z|\theta) = p(x|\theta)p(z|x, \theta)$ e, portanto,

$$p(x|\theta) = \frac{p(x, z|\theta)}{p(z|x, \theta)},$$

sendo a função de log verossimilhança expressa por (Migon & Gamerman, 1999)

$$L(\theta; x) = \log p(x, z|\theta) - \log p(z|x, \theta).$$

Calculando-se os valores esperados com respeito a $p(z|x, \theta)$, como uma forma de eliminar o vetor de quantidades não observadas (Z), obtém-se

$$L(\theta; x) = Q(\theta, \theta^{(0)}) - H(\theta, \theta^{(0)}),$$

onde $\theta^{(0)}$ é um valor inicial atribuído a θ; $Q(\theta, \theta^{(0)}) = E_{Z|X,\theta^{(0)}}[\log p(X, Z|\theta)]$; e $H(\theta, \theta^{(0)}) = E_{Z|X,\theta^{(0)}}[\log p(Z|X, \theta)]$ (Migon & Gamerman, 1999).

O algoritmo EM é, então, implementado iterativamente seguindo os passos abaixo, onde $\theta^{(j)}$ representa o valor de θ na iteração j:
1) passo-E (esperança), através do cálculo de $Q(\theta, \theta^{(j-1)}) = E_{Z|X,\theta^{(j-1)}}[\log l(\theta; Y)]$;
2) passo-M (maximização), que visa encontrar o valor de θ de modo a maximizar a esperança calculada no passo-E.

A sequência $\theta^{(j)}$, $j \geq 1$, encontrada é tal que $L(\theta^{(j)}|x) \leq L(\theta^{(j+1)}|x)$, sendo monotonicamente crescente na verossimilhança $l(\theta; x)$ (Dempster et al., 1977; apud Migon & Gamerman, 1999).

A convergência do algoritmo é obtida adotando-se valores muito pequenos, δ ou ϵ, que satisfaçam, respectivamente, a $|\theta^{(j)} - \theta^{(j-1)}| < \delta$ ou $|Q(\theta^{(j)}, \theta^{(j-1)}) - Q(\theta^{(j-1)}, \theta^{(j-1)})| < \epsilon$ (Migon & Gamerman, 1999).

2.4.2.1 Modelos de Misturas

Seja $y = (y_1, \ldots, y_n)$ uma amostra independente e identicamente distribuída de uma mistura de modelos com densidade

$$p_\theta(u) = \sum_{j=1}^{J} \pi_j \, p_j(u|\theta_j),$$

onde π_j são probabilidades desconhecidas associadas ao modelo de misturas tal que $\sum_j \pi_j = 1$; $p_j(u|\theta_j)$ são modelos probabilísticos; e θ_j são parâmetros desconhecidos.

Desta forma, θ é uma coleção de π_j e de θ_j. Supõe-se, assim, que y_i venha de uma das J populações, mas não se sabe qual delas.

A log-verossimilhança, com base no dado observado y, é expressa por (Pawitan, 2001)

$$L(\theta; y) = \sum_i \log \left\{ \sum_{j=1}^{J} \pi_j \, p_j(y_i|\theta_j) \right\}.$$

Os modelos de misturas são, geralmente, usados em modelos paramétricos ou modelos baseados na metodologia de *cluster*. O principal objetivo é classificar o dado y observado em diferentes *clusters* ou subpopulações (Pawitan, 2001).

Um modelo de misturas, para $J = 2$, pode ser implementado gerando-se, inicialmente, uma amostra obtida através da soma de duas subamostras normalmente distribuídas, a seguir (Pawitan, 2001):

$$\pi_1 \, N(\mu_1, \sigma_1^2) + \pi_2 \, N(\mu_2, \sigma_2^2),$$

onde $\pi_1 + \pi_2 = 1$; e $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$.

A função de log verossimilhança é obtida através de

$$L(\theta; y) = \sum_i \log \left\{ \sum_{j=1}^{J} \pi_1 \phi(y_i, \mu_1, \sigma_1^2) + (1 - \pi_1)\phi(y_i, \mu_2, \sigma_2^2) \right\},$$

onde $\phi(y, \mu, \sigma^2)$ é a densidade da distribuição $N(\mu, \sigma^2)$.

As probabilidades podem ser atualizadas através da seguinte expressão:

$$\pi_j^1 = \frac{\sum_i \widehat{p_{ij}}}{n},$$

onde $\widehat{p_{ij}}$ é a probabilidade estimada de que y_i venha da população j.

A verossimilhança ponderada pelo j-ésimo parâmetro $\theta_j = (\mu_j, \sigma_j)$ será, então

$$-\frac{1}{2} \sum_i \widehat{p_{ij}} \left\{ \log \, \sigma_j^2 + \frac{(y_i - \mu_j)^2}{\sigma_j^2} \right\},$$

de onde se obtém as seguintes atualizações para os parâmetros:

$$\mu_j^1 = \frac{\sum_i \widehat{p_{ij}}\, y_i}{\sum_i \widehat{p_{ij}}} \qquad \text{e} \qquad \sigma_j^{2(1)} = \frac{\sum_i \widehat{p_{ij}}\, (y_i - \mu_j^1)^2}{\sum_i \widehat{p_{ij}}}.$$

2.4.3 Critério de Informação de Akaike

O critério de informação de Akaike (AIC) (Akaike, 1973; 1974) trouxe uma grande contribuição para a seleção de modelos. Akaike usou a informação de Kullback-Leibler (K-L) como uma base fundamental para a escolha de modelos, reconhecendo que os seus parâmetros devem ser calculados a partir dos dados e que existe uma incerteza considerável nessa estimativa (Anderson, 2008).

A descoberta de uma relação entre a informação K-L e a função de máxima verossimilhança tem permitido grandes avanços práticos e teóricos na escolha de modelos e na análise de dados complexos. Segundo deLeeuw (1992; apud Anderson, 2008), Akaike encontrou uma relação formal entre a entropia de Boltzmann e informação K-L (paradigmas dominantes na teoria da informação e codificação) e máxima verossimilhança (cujo paradigma dominante é a estatística).

A informação K-L, $I(f, g)$, representa a perda de informação decorrente do uso de um modelo g para aproximar o modelo verdadeiro f, podendo ser considerada uma medida de discrepância. Essa medida de informação visa encontrar um modelo, dentre um conjunto de modelos candidatos, tal que o valor de $I(f, g)$ seja mínimo (Anderson, 2008).

Akaike mostrou que o desafio, então, consiste na estimação de

$$E_y E_x[\log(g(x|\hat{\theta}(y)))] = \log(l(\hat{\theta}|dados)) - K.$$

Uma associação importante pode ser resumida através da Figura 2.10 (Anderson, 2008):

Entropia negativa	=	Informação K-L & E[Informação K-L]	=	$\log(l) - K$
↓		↓		↓
Termodinâmica		Teoria da informação		Estatística

Figura 2.10: Conexão entre Estatística, Teoria da informação e Informação K-L

O critério AIC é definido, assim, como (Anderson, 2008)

$$AIC = -2\log\Big(L(\hat{\theta}|y)\Big) - 2K,$$

onde K é o número de parâmetros do modelo.

Um critério de informação pode ser usado para a ordenação de modelos com a finalidade de identificar o melhor dentre um conjunto de modelos candidatos. Vale observar que um critério de informação não representa um "teste", não valendo portanto, a associação de conceitos tais como poder do teste, P-valores ou niveis de significância α. Assim, deve-se evitar associar a palavra "significante" aos resultados de um paradigma teórico de informação (Burnham & Anderson, 2002).

Os testes de hipóteses representam um paradigma bem diferente, e geralmente inferior, para a análise de dados complexos. Questões sobre a força da evidência para modelos podem ser melhor analisadas com o uso da razão de evidências, análise dos resíduos R_2 ajustados, além de outros modelos diagnósticos e estatísticas descritivas (Burnham & Anderson, 2002).

2.5 Redes Neurais Artificiais

O cérebro biológico não possui a habilidade de realizar cálculos sucessivos tão rapidamente quanto o computador. No entanto, em algumas situações, o cérebro biológico é capaz de resumir, processar e generalizar dados imediatamente, enquanto um computador convencional precisaria de muito tempo de processamento e, na maioria dos casos, não teria capacidade de entender o problema.

As redes neurais artificiais (RNA) tentam superar as desvantagens dos computadores convencionais, inspirando-se na abstração de processos observados no sistema nervoso biológico humano (Detienne et al., 2003; Kovács, 2002). As RNA emulam o comportamento do cérebro humano, através de modelos computacionais inspirados no neurônio biológico. Frequentemente, as RNA são identificadas como uma subespecialidade de Inteligência Artificial, outras vezes como uma classe de modelos matemáticos para problemas de classificação e reconhecimento de padrões, outras ainda como uma parte da teoria conexionista dos processos mentais e finalmente, como uma categoria de modelos em ciência da cognição (Kovács, 2002).

Segundo Haykin (2001), "uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso".

As RNA podem aprender através da experiência, generalizar com base em exemplos anteriores e abstrair características relevantes por intermédio dos dados de entrada (Wasserman, 1989). A rede neural assemelha-se ao cérebro humano em dois aspectos (Haykin, 2001): (1) o conhecimento é adquirido pela rede, a partir de seu ambiente, através de um processo de aprendizagem; e (2) os pesos sinápticos, forças de conexão entre neurônios (unidades de processamento), são utilizados para armazenar o conhecimento adquirido.

Dentre as diversas aplicações, as RNA podem ser utilizadas para o reconhecimento de fala; localização de fontes de radar; otimização de processos químicos; reconhecimento automático de caracteres escritos à mão; reconhecimento de padrões; diagnóstico médico; diagnóstico de falhas; análise de crédito; classificação; detecção de sinais; e controle de sistema de navegação de aeronaves (Pereira & Rodrigues, 1998; Pereira & Rao, 2009; Braga et al., 2000; Haykin, 2001; Cheng & Titterington, 1994).

A arquitetura da RNA é definida através da organização dos neurônios e dos tipos de conexões permitidas, classificando-se em: (1) rede alimentada adiante (*feedforward*) ou acíclica, quando a saída de um neurônio não é usada como entrada de nodos em camadas anteriores, ou seja não há realimentação; e (2) rede do tipo *feedback* ou cíclica, quando a saída de um neurônio da i-ésima camada é usada como entrada de nós em camadas de índice menor ou igual a i (Pereira & Rao, 2009; Braga et al., 2000).

2.5.1 Redes Neurais do Tipo *Feedforward*

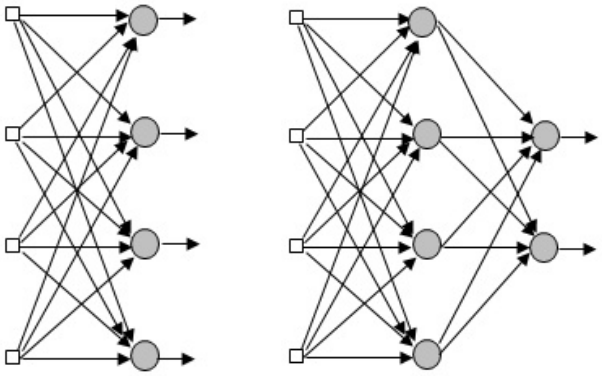
Uma rede neural do tipo *feedforward* é definida por um grafo orientado acíclico e por uma escolha de funções utilizadas em seus nós (Fine, 1999), caracterizando-se por dois tipos de arquitetura (Haykin, 2001):
1) RNA com camada única, segundo o modelo de McCulloch & Pitts (1943), constituída de um único neurônio, unidade fundamental de processamento de uma rede neural. Esses tipos de redes neurais conseguem resolver, apenas, problemas linearmente separáveis (Braga et al., 2000).
2) RNA com múltiplas camadas (*multilayer perceptron* - MLP), caracterizadas pela presença de uma ou mais camadas escondidas, onde os neurônios conectam-se apenas com os neurônios das camadas posteriores, não havendo conexões entre neurônios pertencentes a uma mesma camada.

A adição de uma ou mais camadas escondidas possibilita a extração de características de ordem elevada. Entretanto, uma única camada oculta é suficiente para um *perceptron* de múltiplas camadas computar uma aproximação uniforme para um dado conjunto de treinamento e a saída desejada (Haykin, 2001).

A Figura 2.11(a) (Haykin, 2001) apresenta a rede completamente conectada sem realimentação com camada única, a camada de saída de nós computacionais (neurônios). Uma rede é totalmente conectada quando cada nó de uma camada encontra-se ligado a cada um dos nós da camada seguinte.

Por convenção, adota-se a nomenclatura $X - Y - Z$ para representar a topologia de uma rede neural constituida de X nós na entrada, Y neurônios na camada intermediária e Z neurônios na saída da rede neural. A Figura 2.11(b) (Haykin, 2001)

representa uma rede 4-4-2, possuindo 4 nós de entrada, 4 neurônios na camada intermediária e 2 na saída.



(a) Camada única (b) Topologia 10-4-2

Figura 2.11: Redes neurais do tipo *feedforward*

A aprendizagem pode ser do tipo supervisionado ou não supervisionado. A aprendizagem supervisionada é aquela na qual, para cada vetor de entrada, obtém-se uma saída conhecida, sendo possível então, modificar os pesos sinápticos de maneira ordenada com a finalidade de atingir o objetivo desejado.

O aprendizado não supervisionado caracteriza-se pelo desconhecimento da saída, para cada vetor de entrada. A rede desenvolve a habilidade de formar representações internas para codificar as características de entrada e criar novas classes ou grupos automaticamente. Isso é possível, apenas, se houver redundância nos dados de entrada, sem a qual não se pode encontrar quaisquer padrões ou características dos dados de entrada (Braga et al., 2000). O treinamento não supervisionado equipara-se aos métodos estatísticos de análise de agrupamento e de componentes principais (Pereira & Rodrigues, 1998; Pereira & Rao, 2009).

O processo de aprendizagem envolve a modificação iterativa dos pesos até que seja minimizado o erro calculado entre a saída desejada e a obtida pela rede. Dessa forma, a rede neural é capaz de extrair as informações relevantes, inerentes aos dados.

A fase de treinamento tem como objetivo, alcançar um equilíbrio entre a capacidade de responder corretamente aos padrões de entrada usados no treinamento (memorização) e a habilidade para dar respostas razoáveis às entradas similares, mas não idênticas, as usadas no treinamento (generalização) (Fine, 1999).

O treinamento das redes neurais do tipo MLP pode ser realizado com o uso do algoritmo de retropropagação (*backpropagation*), através da soma ponderada das entradas e do ajuste dos pesos, considerando a diferença entre a resposta desejada e a obtida pela rede. Geralmente, adota-se, como medida de desempenho da rede

neural, o erro quadrático médio ou a soma dos erros quadrados, calculados sobre a amostra de treino (Haykin, 2001).

2.5.1.1 Regularização Bayesiana

Além de buscar o menor erro no conjunto de treinamento, o modelo neural visa à obtenção da melhor resposta para novos exemplos (capacidade de predição). O método de regularização consiste em restringir o tamanho dos pesos da rede para melhorar a sua capacidade de predição, objetivo esse que é alcançado reduzindo-se os valores dos pesos, obtendo-se assim um grau de suavização na resposta da rede (Foresee & Hagan, 1997).

As abordagens tradicionais para a seleção de pesos estão fortemente relacionadas à estimação por máxima verossimilhança. Assim, é também comum o uso da abordagem Bayesiana (Cheng & Titterington, 1994). Da inferência Bayesiana, o teorema de Bayes é expresso por (Migon & Gamerman, 1999):

$$p(\theta|x,H) = \frac{p(\theta,x|H)}{p(x|H)} = \frac{p(x|\theta,H)p(\theta|H)}{p(x|H)},$$

sendo $p(x|H) = \int_{\Theta} p(x,\theta|H)d\theta$. Esse teorema pode ser reescrito como

$$p(\theta|x) \propto p(x|\theta)p(\theta),$$

onde θ é um parâmetro populacional desconhecido, que se deseja estimar; $p(\theta)$ representa a probabilidade a priori para θ; $p(\theta|x)$ é a probabilidade a posteriori de θ, dado o conjunto x de dados e; $p(x|\theta)$ é a verossimilhança, uma medida de plausibilidade para θ.

Em geral, a função objetivo do modelo neural visa à minimização da soma dos quadrados dos erros, ou seja, minimizar

$$F = E_D = \sum_{i=1}^{n}(t_i - a_i)^2,$$

onde a_i é a saída da rede neural e t_i representa a resposta desejada. A regularização é obtida através da adição de um termo à função objetivo, tornando-se

$$F = \beta E_D + \alpha E_W,$$

onde E_W é a soma dos quadrados dos pesos; α e β representam os parâmetros de regularização. Se $\alpha \ll \beta$, o algoritmo fornecerá erros menores. Entretanto, se $\alpha \gg \beta$, a redução dos pesos da rede receberá maior ponderação, produzindo assim uma resposta mais suave para a rede (Foresee & Hagan, 1997).

Na abordagem Bayesiana, os pesos da rede são variáveis aleatórias que podem ser

atualizados através do cálculo da posteriori (Mackay, 1992; apud Foresee & Hagan, 1997)

$$p(w|D,\alpha,\beta,M) = \frac{p(D|w,\beta,M)p(w|\alpha,M)}{p(D|\alpha,\beta,M)}, \tag{2.19}$$

onde D representa o conjunto de dados; M é o modelo neural utilizado; e w é o vetor de pesos; $P(w|\alpha,M)$ é a probabilidade a priori; $P(D|w,\beta,M)$ é a verossimilhança. O denominador da equação 2.19 representa o fator de normalização que, segundo a regra da probabilidade total, deve somar 1.

Os pesos ótimos devem maximizar a probabilidade a posteriori $P(w|D,\alpha,\beta,M)$, que é equivalente a minimizar a função objetivo regularizada $F = \beta E_D + \alpha E_W$ (Mackay, 1992; apud Foresee & Hagan, 1997).

Entretanto, um grande desafio encontrado para a implementação da Regularização Bayesiana está relacionado à escolha dos parâmetros α e β da função objetivo, que podem ser obtidos através de

$$p(\alpha,\beta|D,M) = \frac{p(D|\alpha,\beta,M)p(\alpha,\beta|M)}{p(D|M)},$$

que envolve o cálculo da matriz Hessiana, cuja implementação exige um grande esforço computacional. Para contornar esse problema, Foresee & Hagan (1997) propuseram o uso da aproximação de Gauss-Newton, através do algoritmo de Levenberg-Marquardt para treinamento da rede neural.

2.5.2 Mapas Auto-Organizáveis de Kohonen

Os Mapas Auto-Organizáveis (*Self-Organizing Map*) - SOM (Kohonen, 1982; Kohonen, 1990), ou mapas de Kohonen, constituem uma classe especial de redes neurais, tendo como base a aprendizagem competitiva, sendo eficazes para a visualização de dados de grandes dimensões.

Os mapas SOM podem ser definidos como um mapeamento não linear, ordenado e suavizado, dos dados de entrada de grande dimensão, em uma matriz regular de baixa dimensão (Kohonen, 2001). Nesse mapeamento, os dados de entrada são projetados numa grade de nós discretos interconectados, cada nó representando um agrupamento ou *cluster* homogêneo de pontos (Izenman, 2008).

Basicamente, o algoritmo produz um grafo de similaridades, convertendo relações estatísticas não lineares, que envolvem dados de grandes dimensões, em relações geométricas simples, através da projeção dos seus pontos numa grade (rede) de baixa dimensão. Em geral, os neurônios dessa rede são dispostos numa grade unidimensional ou bidimensional (Kohonen, 2001; Haykin, 2001). Desta forma, os mapas de Kohonen comprimem a informação em uma rede, partindo do espaço de pontos

original a fim de obter um conjunto de pontos bem menor, sendo este representativo e tendo como propriedade manter preservadas as relações de distância e vizinhança (Kohonen, 2001; Pereira & Rao, 2009).

Os neurônios vencedores são ordenados entre si, criando sobre essa grade um sistema de coordenadas para cada característica de entrada (Haykin, 2001).

Após um grande número de iterações, o algoritmo SOM permite uma visualização gráfica do espaço de saída, que consiste numa grade (ou rede) de nós interconectados, os neurônios artificiais. Para uma saída bidimensional, os nós são tipicamente organizados numa grade composta por formas quadradas, retangulares ou hexagonais interligadas. Para melhor visualização, as formas hexagonais são preferíveis, conforme mostra a Figura 2.12 (Izenman, 2008).

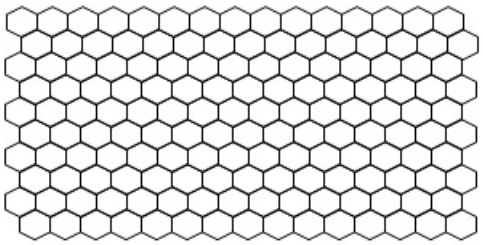


Figura 2.12: Mapa auto-organizável hexagonal de dimensão 10×15

O número total de nós K é, em geral, definido empiricamente, atribuindo-se, a princípio, um valor bem maior que o número suposto de *clusters* dos dados. Nas etapas seguintes, o valor de K é reduzido gradativamente (Izenman, 2008).

Para a implementação do algoritmo SOM, define-se inicialmente a dimensão do mapa através da escolha de sua altura K_1 e da largura K_2. No passo seguinte, inicializam-se aleatoriamente todos os pesos m_k. Em seguida, dá-se início ao processo iterativo, obedecendo as etapas abaixo (Izenman, 2008):

1) Seleciona-se, aleatoriamente, uma variável de entrada, fazendo a sua padronização de forma a apresentar média zero e variância um. Este procedimento evitará a influência indevida dessa variável sobre o resultado.

2) Calcula-se a distância Euclidiana entre a variável padronizada X e cada m_k, identificando o nó que apresenta a menor distância com relação a X.

3) Considera-se como "nó vencedor", relativamente a X, ou "*best-matching unit*" (BMU), o nó k^* tal que

$$k^* = \arg\min_k \{||X, m_k||\},$$

onde $||\,\cdot\,||$ representa a norma Euclidiana.

4) Define-se o nó $k' \in \mathcal{K}$ como um nó vizinho do nó $k \in \mathcal{K}$, se a distância Euclidiana entre os pesos associados a esses nós, m_k e $m_{k'}$, for inferior a um limiar c, escolhido arbitrariamente. O conjunto de nós vizinhos $\mathcal{N}_c(K^*)$ é, então, considerado a vizinhança do nó vencedor k^*.

5) Atualiza-se o nó k^* e todos os seus vizinhos k', de forma tal que cada peso m_k, $k \in \mathcal{N}_c(K^*)$, esteja mais próximo de X. Os pesos podem ser atualizados através do cálculo de

$$\Delta m_k = \begin{cases} \alpha(X - m_k), & \text{se } k \in \mathcal{N}_c(K^*) \\ 0, & \text{caso contrário,} \end{cases}$$

onde α é a taxa de aprendizagem, $0 < \alpha < 1$. Assim, m_k passa a assumir o valor $m_k + \Delta m_k$, se $k \in \mathcal{N}_c(K^*)$, permanecendo inalterado em caso contrário.

As etapas do algoritmo são repetidas, iterativamente, um grande número de vezes, percorrendo o conjunto de vetores de entrada, um de cada vez. Para que se obtenha uma boa acurácia, Kohonen (2001) recomenda que o número de iterações seja, pelo menos, 500 vezes o número de neurônios do mapa.

Uma boa precisão estatística pode ser alcançada (1) fixando-se a taxa de aprendizagem em um valor pequeno ($\leq 0,01$), durante a convergência do algoritmo, para um número de iterações relativamente grande (em torno de 1000); e (2) mantendo-se na região de vizinhança somente os vizinhos mais próximos do neurônio vencedor, no início da fase de convergência, podendo ser reduzido a um ou zero neurônios (Haykin, 2001).

O conceito de distância ponderada tem sido uma estratégia bastante utilizada, na qual a atualização dos pesos se dá por intermédio do cálculo de (Izenman, 2008)

$$\Delta m_k = \begin{cases} \alpha h_k(X - m_k), & \text{se } k \in \mathcal{N}_c(K^*) \\ 0, & \text{caso contrário,} \end{cases}$$

onde h é uma função de vizinhança que depende de quão próximo os vizinhos se encontram de m_{k^*}. Quanto maior a proximidade, melhor o ajuste.

A função h_k é inversamente proporcional à distância, assumindo o valor 1 quando a distância for igual a zero e decrescendo à medida que as distâncias se tornem maiores. Para $k \notin \mathcal{N}_c(K^*)$, atribui-se a h_k o valor zero.

A função de vizinhança, mais amplamente utilizada, é a função *kernel* Gaussiana multivariada, expressa por (Izenman, 2008)

$$h_k = \exp\left\{ - \frac{||m_k - m_{k^*}||^2}{2\sigma^2} \right\} I_{[k \in \mathcal{N}_c(K^*)]},$$

onde $\sigma > 0$ representa o raio de vizinhança.

Embora α, c e σ sejam escolhidos arbitrariamente, seus valores poderão ser modificados durante o processo iterativo. Por exemplo, o valor limiar c pode ser reduzido a 1 durante as 1000 primeiras iterações ou a partir de um valor C, também estipulado arbitrariamente.

A taxa de aprendizagem, $0 < \alpha < 1$, assume valores decrescentes e pode ser

calculada segundo as formas abaixo (Izenman, 2008), sendo uma função da época atual t:

1) Linear: $\alpha(t) = \alpha_0(1 - t/T)$;

2) Poder: $\alpha(t) = \alpha_0(0,005/\alpha_0)^{t/T}$;

3) Inversa: $\alpha(t) = \alpha_0/(1 + 100t/T)$,

onde α_0 é a taxa de aprendizagem inicial; e T representa o número total de iterações.

Os mapas de Kohonem podem ser utilizados na solução de problemas de clusterização envolvendo diversas áreas tais como sistemas de informação geográfica (GIS), bioinformática, investigação médica, antropologia física, processamento de linguagem natural, sistemas de recuperação de documento e ecologia (Izenman, 2008). Os mapas SOM, podem ser considerados uma generalização não-linear da análise de componentes principais (Ritter, 1995 apud Haykin, 2001).

Capítulo 3

Discussão e Conclusões

Uma das atribuições da Divisão de Pesquisa (DPq) do Hospital Universitário Clementino Fraga Filho - HUCFF/UFRJ consiste na prestação de consultoria estatística em projetos de pesquisa clínica e básica em saúde, bem como na área de administração hospitalar. Projetos estes, onde é necessária a utilização de métodos quantitativos e a habilidade no uso de *softwares* estatísticos. A necessidade de apoio especializado tornou-se evidente através do grande número de pesquisadores do HUCFF, que se dirige à DPq em busca de auxílio relacionado às análises estatísticas de suas teses e/ou artigos. O atendimento da demanda de pesquisadores do HUCFF é viabilizado através de consultorias realizadas pelo Núcleo de Assessoria Estatística da DPq, para a elaboração de modelos estatísticos e suas respectivas análises.

Para a realização dessas consultorias, a DPq conta com a participação de alunos de pós-graduação da COPPE, além de alunos de graduação do Instituto de Matemática - IM/UFRJ, através dos programas de iniciação científica. O trabalho no Núcleo de Assessoria Estatística da DPq propicia um aprendizado aos estudantes (graduação e pós-graduação) que, além da prática estatística, aprendem a interagir com outros profissionais, no caso, os pesquisadores clínicos.

Este trabalho apresentou diferentes técnicas de aprendizado estatístico com diversas aplicações na pesquisa médica, frutos de diversas consultorias desenvolvidas na DPq/HUCFF. Aplicações estas que deram origem a oito artigos já publicados, aceitos ou submetidos.

A técnica das árvores de sobrevida (Seção 2.3.3) foi utilizada conjuntamente com o MELD para descrever a mortalidade na fila de espera por transplante de fígado no longo prazo (Apêndice B). Os modelos estruturados em árvore permitiram a identificação dos grupos de risco através da visualização gráfica.

O uso não convencional das curvas de sobrevida (Seção 2.2) foi apresentado através da aplicação relacionada à síndrome do olho seco, em pacientes infectados pelo vírus HIV (Apêndice C), e na identificação de alterações gustatórias em pacientes

com otite média crônica (Apêndice G). O método foi proposto como uma alternativa para os testes estatísticos tradicionais para a comparação de amostras compostas por variáveis qualitativas ordinais.

O Algoritmo EM (Seção 2.4.2) e o critério de Akaike (Seção 2.4.3) foram utilizados para descrever a distribuição da antigenicidade de plaquetas em indívíduos euro e afro-brasileiros do subgrupo sanguíneo A (Apêndice D), através da escolha de modelos de misturas de normais.

O estudo apresentado no Apêndice F busca determinar a relação entre alteração contrátil e arritmias ventriculares complexas em pacientes chagásicos. A análise dos dados foi realizada através da implementação dos modelos log-lineares (Seção 2.1.3). O uso do modelo log-linear gráfico (Seção 2.1.3.3) permitiu que fossem, visualmente, identificadas as relações de independência condicional entre as variáveis envolvidas no modelo log-linear.

As árvores de regressão (Seção 2.3.2) e as curvas de sobrevida, para análise de concordância entre medidas (*Survival Agreement Plot*), foram utilizadas para avaliar a função do sistema cardiovascular autônomo através da cintilografia com MIBG-Iodo-123 em pacientes chagásicos (Apêndice H).

As redes neurais artificiais do tipo *feedforward* (Seção 2.5.1) foram utilizadas para investigar a associação entre a poluição atmosférica e condições climáticas no número de internações hospitalares, motivadas por bronquiolite infantil (Apêndice E). Uma revisão dos estudos e metodologias utilizadas nesse tipo de investigação é apresentada no Apêndice I.

Apesar de não relacionado à área médica, o nono artigo, apresentado no Apêndice A, foi incluído para efeito ilustrativo do uso das técnicas abaixo, cuja utilização em conjunto constitui um ferramental de grande utilidade na pesquisa clínica:

1. Mapas auto-organizáveis de Kohonen (Seção 2.5.2), para a solução de problemas de clusterização, tais como: determinar o desfecho e combinar receptores e doadores no transplante de fígado (Haydon et al., 2005); reconstrução do córtex cerebral (Chuang et al., 2007); reconhecimento de padrões em magnetocardiografias (Naenna et al., 2004); avaliação dos fatores relacionados à síndrome da resistência à insulina e doenças cardiovasculares (Valkonen et al., 2002); e seleção de micromatrizes de genes para predição de câncer (Vanichayobon et al., 2007);

2. Rede neural do tipo MLP (Seção 2.5.1), para problemas de classificação, usando a clusterização obtida pelo SOM como saída desejada para a rede. Adicionalmente, a implementação do modelo MLP permite a identificação das variáveis de maior relevância;

3. Análise de componentes principais (Seção 2.4.1) para redução da dimensionalidade dos dados, onde as variáveis são substituídas por um número menor de funções

lineares dessas variáveis; e

4. Gráfico de coordenadas paralelas, para representação de dados multivariados e consequente visualização de possível clusterização.

Referências Bibliográficas

AALEN, O. O. "Non Parametric Inference for a Family of Counting Processes", *Annals of Statistics*, v. 6, n. 4, pp. 701–726, 1978.

AALEN, O. O., BORGAN, O., GJESSING, H. K. *Survival and Event History Analysis: A Process Point of View*. Springer, 2008.

AGRESTI, A. *Categorical Data Analysis*. 2 ed. New Jersey, Wiley, 2002.

AGRESTI, A. *An Introduction to Categorical Data Analysis*. 2 ed. New Jersey, Wiley, 2007.

AKAIKE, H. "Information Theory and an Extension of the Maximum Likelihood Principle". In: Petrov, B. N., Csaki, F. (Eds.), *Second International Symposium on Information Theory*, pp. 267–281, Budapest, 1973.

AKAIKE, H. "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control*, v. 19, pp. 716–723, 1974.

ALTMAN, D. G., BLAND, J. M. "Improving doctors's understanding of statistics", *Journal of the Royal Statistical Society A*, v. 154, pp. 223–267, 1991.

ANDERSON, D. R. *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer, 2008.

BERK, R. A. "Data Mining within a Regression Framework". In: *The Data Mining and Knowledge Discovery Handbook*, pp. 231–255, 2005.

BERK, R. A. *Statistical Learning from a Regression Perspective*. Springer, 2008.

BERKSON, J., GAGE, R. P. "Survival Curve for Cancer Patients Following Treatment", *Journal of the American Statistical Association*, v. 47, n. 259, pp. 501–515, 1952.

BOAG, J. W. "Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy (with discussion)", *Journal of the Royal Statistical Society B*, v. 11, n. 1, pp. 15–53, 1949.

BOAVENTURA-NETTO, P. O. *Grafos: Teoria, Modelos, Algoritmos*. 4 ed. São Paulo, Edgard Blücher, 2006.

BOHORIS, G. "Comparison of the Cumulative-Hazard and Kaplan-Meier Estimators of the Survivor Function", *IEEE Transactions on Reliability*, v. 43, n. 2, pp. 230–232, 1994.

BRAGA, A. P., CARVALHO, A. P. L. F., LUDERMIR, T. B. *Redes Neurais Artificiais: Teoria e Aplicações*. LTC Editora, 2000.

BREIMAN, L. "Statistical Modeling: The Two Cultures", *Statistical Science*, v. 16, n. 3, pp. 199–215, 2001.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., et al. *Classification and Regression Trees*. California, Wadsworth, 1984.

BRESLOW, N. "A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship", *Biometrika*, v. 57, n. 3, pp. 579–594, 1970.

BURNHAM, K. P., ANDERSON, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2002.

CARVALHO, M. S., ANDREOZZI, V. L., CODEÇO, C. T., et al. *Análise de Sobrevida: Teoria e Aplicações em Saúde*. Editora Fiocruz, 2005.

CHATFIELD, C. "Model Uncertainty, Data Mining and Statistical Inference", *Journal of the Royal Statistical Society A*, v. 158, n. 3, pp. 419–466, 1995.

CHENG, B., TITTERINGTON, D. M. "Neural Networks: A Review from a Statistical Perspective", *Statistical Science*, v. 9, n. 1, pp. 2–30, 1994.

CHUANG, C.-H., CHENG, P. E., LIOU, M., et al. "Application of self-organizing map (SOM) for cerebral cortex reconstruction", *International Journal of Computational Intelligence Research*, v. 3, n. 1, pp. 26–30, 2007.

CIAMPI, A., ETEZADI-AMOLI, J. "A General Model for Testing the Proportional Hazards and the Accelerated Failure Time Hypotheses in the Analysis of Censored Survival Data with Covariates", *Communications in Statistics*, v. 14, n. 3, pp. 651–667, 1985.

CIAMPI, A., THIFFAULT, J., NAKACHE, J.-P., et al. "Stratification by Stepwise Regression, Correspondence Analysis and Recursive Partition: A Comparison of Three Methods of Analysis for Survival Data with Covariates", *Computational Statistics and Data Analysis*, v. 4, n. 3, pp. 185–204, 1986.

COLOSIMO, E. A., GIOLO, S. R. *Análise de Sobrevivência Aplicada*. Edgard Blücher, 2006.

COPPI, R. "A Theoretical Framework for Data Mining: The Informational Paradigm", *Computational Statistics and Data Analysis*, v. 38, pp. 501–515, 2002.

COX, D. R. "Regression Models and Life-Tables (with discussion)", *Journal of the Royal Statistical Society B*, v. 34, n. 2, pp. 187–220, 1972.

COX, D. R., OAKES, D. *Analysis of Survival Data*. Chapman and Hall, 1984.

CRAMER, J. S. *Logit Models from Economics and Other Fields*. Cambridge University Press, 2003.

CROWLEY, J., LEBLANC, M., GENTLEMAN, R., et al. "Exploratory Methods in Survival Analysis", *Lecture Notes-Monograph Series*, v. 27, pp. 55–77, 1995.

DARROCH, J. N., LAURITZEN, S. L., SPEED, T. P. "Markov Fields and Log-Linear Interaction Models for Contingency Tables", *The Annals of Statistics*, v. 8, n. 3, pp. 522–539, 1980.

DAWID, A. P. "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society B*, v. 41, n. 1, pp. 1–31, 1979.

deLEEUW, J. "Breakthroughs in statistics". v. 1, cap. Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle, pp. 599–609, Springer-Verlag, 1992.

DEMÉTRIO, C. G. B. "Modelos Lineares Generalizados em Experimentação Agronômica". , 2002. Disponível em: `http://ce.esalq.usp.br/clarice/Apostila.pdf`, acesso em 05 jun. 2010.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society B*, v. 39, n. 1, pp. 1–38, 1977.

DETIENNE, K. B., DETIENNE, D. H., JOSHI, S. A. "Neural Networks as Statistical Tools for Business Researchers", *Organizational Research Methods*, v. 6, n. 2, pp. 236–265, 2003.

DOBSON, A. J. *An Introduction to Generalized Linear Models*. 2 ed. , Chapman and Hall, 2002.

DOWNS, S. M., WALLACE, M. Y. "Mining Association Rules from a Pediatric Primary Care Decision Support System". In: Overhage, J. M. (Ed.), *Converging Information, Technology, and Health Care*, pp. 200–204, Los Angeles, Ca, 2000. AMIA Annual Symposium.

EDWARDS, D. *Introduction to Graphical Modelling*. New York, Springer, 2000.

FELLER, W. *An Introduction to Probability Theory and its Applications*, v. 1. Wiley, 1968.

FINE, T. L. *Feedforward Neural Network Methodology*. New York, Springer, 1999.

FORESEE, F. D., HAGAN, M. T. "Gauss-Newton Approximation to Bayesian Learning". In: *IEEE International Conference on Neural Networks*, pp. 1930–1935, Houston, TX, USA, 1997.

FRAWLEY, W., PIATETSKY-SHAPIRO, G., MATHEUS, C. "Knowledge Discovery in Databases: An Overview", *AI Magazine*, v. 13, n. 3, pp. 57–70, 1992.

FRIEDMAN, J. H. "Data Mining and Statistics: What's the Connection?" . 29th Symposium on the Interface: Computing Science and Statistics, 1997.

GEHAN, E. A. "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples", *Biometrika*, v. 52, pp. 203–223, 1965.

GORDON, L., OLSHEN, R. A. "Tree-Structured Survival Analysis", *Cancer Treatment Reports*, v. 69, n. 10, pp. 1065–1069, 1985.

GREEN, S., BENEDETTI, J., CROWLEY, J. *Clinical trials in oncology*. Chapman & Hall/CRC, 2003.

HAIR, J. F., TATHAM, R. L., ANDERSON, R. E., et al. *Multivariate Data Analysis*. 5 ed. New Jersey, Prentice Hall, 1998.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2 ed. New York, Springer, 2009.

HAYDON, G. H., HILTUNEN, Y., LUCEY, M. R., et al. "Self-organizing maps can determine outcome and match recipients and donors at orthotopic liver transplantation", *Transplantation*, v. 79, n. 2, pp. 213–218, 2005.

HAYKIN, S. *Redes Neurais: Princípios e Prática*. 2 ed. Porto Alegre, Bookman, 2001.

HOLMES, J. H., DURBIN, D. R., WINSTON, F. K. "Discovery of Predictive Models in an Injury Surveillance Database: An Application of Data Mining in Clinical Research". In: Overhage, J. M. (Ed.), *Converging Information, Technology, and Health Care*, pp. 359–363, Los Angeles, Ca, 2000. AMIA Annual Symposium.

HOTHORN, T., LAUSEN, B., BENNER, A., et al. "Bagging Survival Trees", *Statistics in Medicine*, v. 23, pp. 77–91, 2004.

IZENMAN, A. J. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, 2008.

JOHNSON, R. A., WICHERN, D. W. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall, 2007.

KALBFLEISCH, J. D., PRENTICE, R. L. *The Statistical Analysis of Failure Time Data*. Wiley, 1980.

KAPLAN, E. L., MEIER, P. "Nonparametric Estimation From Incomplete Observations", *Journal of the American Statistical Association*, v. 53, pp. 457–481, 1958.

KLEINBAUM, D. G., KLEIN, M. *Survival Analysis: A Self-Learning Text*. Springer, 2005.

KOHONEN, T. "Self-Organized Formation of Topologically Correct Feature Maps", *Biological Cybernetics*, v. 43, pp. 59–69, 1982.

KOHONEN, T. "The Self-Organizing Map", *Proceedings of the IEEE*, v. 78, n. 9, pp. 1464–1480, 1990.

KOHONEN, T. *Self-Organizing Maps*. Springer, 2001.

KOVÁCS, Z. L. *Redes Neurais Artificiais: Fundamentos e Aplicações*. 3 ed. São Paulo, Editora Livraria da Física, 2002.

LEE, E. T., GO, O. T. "Survival Analysis in Public Health Research", *Annual Review of Public Health*, v. 18, n. 1, pp. 105–134, 1997.

LOUZADA-NETO, F. "Extended Hazard Regression Model for Reliability and Survival Analysis", *Lifetime Data Analysis*, v. 3, pp. 367–381, 1997.

LOUZADA-NETO, F. "Modelling Lifetime Data: A Graphical Approach", *Applied Stochastic Models in Business and Industry*, v. 15, pp. 123–129, 1999.

LOUZADA-NETO, F., PEREIRA, B. B. "Modelos em Análise de Sobrevivência", *Cadernos de Saúde Coletiva*, v. 8, n. 1, pp. 9–26, 2000.

LOVELL, M. C. "Data Mining", *The Review of Economics and Statistics*, v. 65, n. 1, pp. 1–12, 1983.

MACHIN, D., CHEUNG, Y. B., PARMAR, M. K. B. *Survival Analysis: A Practical Approach*. Wiley, 2006.

MACKAY, D. J. C. "Bayesian Interpolation", *Neural Computation*, v. 4, pp. 415–447, 1992.

MANLY, B. F. J. *Métodos Estatísticos Multivariados: Uma Introdução*. 3 ed. Porto Alegre, Bookman, 2008.

MANTEL, N. "Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration", *Cancer Chemotherapy Reports*, v. 50, n. 3, pp. 163–170, 1966.

MANTEL, N., HAENSZEL, W. "Statistical Aspects of the Analysis of Data from Retrospective Studies of disease", *Journal of the National Cancer Institute*, v. 22, pp. 719–748, 1959.

MARTINEZ, E. Z. "Contribuições do estatístico para o atendimento ao CONSORT Checklist", *Revista da Faculdade de Medicina de Ribeirão Preto*, v. 42, n. 1, pp. 22–30, 2009.

McCULLAGH, P., NELDER, J. A. *Generalized Linear Models*. London, Chapman and Hall, 1989.

McCULLOCH, W. S., PITTS, W. H. "A Logical Calculus of the Ideas Immanent in Nervous Activity", *Bulletin of Mathematical Biophysics*, v. 5, pp. 115–133, 1943.

MIGON, H. S., GAMERMAN, D. *Statistical Inference: An Integrated Approach*. London, Edward Arnold, 1999.

MONTGOMERY, D. C., PECK, E. A., VINING, G. G. *Introduction to Linear Regression Analysis*. New York, Wiley, 2001.

MULLINS, I. M., SIADATY, M. S., LYMAN, J., et al. "Data Mining and Clinical Data Repositories: Insights from a 667,000 Patient Data Set", *Computers in Biology and Medicine*, v. 36, n. 12, pp. 1351–1377, 2006.

NAENNA, T., EMBRECHTS, M. J., SZYMANSKI, B. "Automated magnetocardiogram classifications with self-organizing maps (SOMs)", *IEEE TENCON*, v. 2, n. 458-460, 2004.

NELDER, J. A., WEDDERBURN, R. W. M. "Generalized Linear Models", *Journal of the Royal Statistical Society A*, v. 135, n. 3, pp. 370–384, 1972.

NELSON, W. "Theory and Applications of Hazard Plotting for Censored Failure Data", *Technometrics*, v. 42, n. 1, pp. 12–25, 2000.

PAWITAN, Y. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, 2001.

PEREIRA, B. B. "Estatística: A Tecnologia da Ciência", *Boletim da Associação Brasileira de Estatística*, v. 13, pp. 27–35, 1997.

PEREIRA, B. B. "Estatística: A Tecnologia da Ciência II", *Boletim da Associação Brasileira de Estatística*, v. 16, n. 47, pp. 37–39, 2000.

PEREIRA, B. B., RAO, C. R. *Data Mining Using Neural Networks: A Guide for Statisticians*. Textbook Revolution, 2009. Disponível em: `http://textbookrevolution.org/index.php/Data_Mining_using_Neural_Networks:A_Guide_for_Statisticians`, acesso em 08 dez. 2009.

PEREIRA, B. B., RODRIGUES, C. V. S. "Redes Neurais em Estatistica". . 13º SINAPE - ABE-Associação Brasieira de Estatística, 144 p., Rio de Janeiro, 1998.

PETO, R., PETO, J. "Asymptotically Efficient Rank Invariant Test Procedures (with Discussion)", *Journal of the Royal Statistical Society A*, v. 135, pp. 185–207, 1972.

POCOCK, S. J. *Clinical trials: A practical approach*. Chichester, Wiley, 1983.

PRATHER, J. C., LOBACH, D. F., GOODWIN, L. K., et al. "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse". In: Masys, D. R. (Ed.), *The Emergence of Internetable Health Care. Systems that Really Work*, pp. 101–105, Nashville, TN, 1997. AMIA Annual Fall Symposium.

PRENTICE, R. L. "Linear Rank Tests with Right Censored Data", *Biometrika*, v. 65, n. 1, pp. 167–179, 1978.

PRICE, D. L., MANATUNGA, A. K. "Modelling Survival Data with a Cured Fraction Using Frailty Models", *Statistics in Medicine*, v. 20, pp. 1515–1527, 2001.

R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. Disponível em: <http://www.R-project.org>. ISBN 3-900051-07-0.

RAO, C. R. *Statistics and Truth: Putting Chance to Work*. Singapore, World Scientific, 1997.

RAO, C. R., TOUTENBURG, H., SHALABH, et al. *Linear Models and Generalizations: Least Squares and Alternatives*. 3 ed. Berlin, Springer, 2008.

RITTER, H. J. *Self-Organizing Feature Maps: Kohonen Maps*. The Handbook of Brain Theory and Neural Networks. The MIT Press, Cambridge, Massachusetts, London, 1995.

RODRIGUES, J., CANCHO, V. G., CASTRO, M. "Teoria Unificada de Análise de Sobrevivência". . 18º SINAPE, 94 p., São Paulo - SP: Associação Brasileira de Estatística, 2008.

SAHAI, H., KHURSHID, A. *Statistics in Epidemiology: Methods, Techniques, and Applications*. Boca Raton, CRC Press, 1996.

SCULLY, K. W., PATES, R. D., DESPER, G. S., et al. "Development of an Enterprise-Wide Clinical Data Repository: Merging Multiple Legacy Databases". In: Masys, D. R. (Ed.), *The Emergence of Internetable Health Care. Systems that Really Work*, pp. 32–36, Nashville, TN, 1997. AMIA Annual Fall Symposium.

SEGAL, M. R. "Regression Trees for Censored Data", *Biometrics*, v. 44, pp. 35–47, 1988.

SHORACK, G. R., WELLNER, J. A. *Empirical Processes with Applications to Statistics*. New York, Wiley, 1986.

STIGLER, S. M. "Gauss and the Invention of Least Squares", *The Annals of Statistics*, v. 9, n. 3, pp. 465–474, 1981.

STIGLER, S. M. *The History of Statistics*. Cambridge, Mass., 1986.

TARONE, R., WARE, J. "On Distribution-free Tests for Equality of Survival Distributions", *Biometrika*, v. 64, pp. 156–160, 1977.

TURA, B. R. *Aplicação do 'Data Mining' em Medicina.* Msc dissertação, NESC - UFRJ, Rio de Janeiro - RJ, 2001.

VALKONEN, V.-P., KOLEHMAINEN, M., LAKKA, H.-M., et al. "Insulin resistance syndrome revisited: Application of self-organizing maps", *International Journal of Epidemiology*, v. 31, 2002.

VANICHAYOBON, S., WICHAIDIT, S., WETTAYAPRASIT, W. "Microarray Gene Selection Using Self-Organizing Map". In: *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 239–244, Beijing, China, 2007.

VAUPEL, J. W., MANTON, K. G., STALLARD, E. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality", *Demography*, v. 16, n. 3, pp. 439–454, 1979.

WASSERMAN, L. *All of Statistics: A Concise Course in Statistical Inference.* New York, Springer, 2004.

WASSERMAN, P. D. *Neural Computing: Theory and Practice.* New York, Van Nostrand Reinhold, 1989.

WHITTAKER, J. *Graphical Models in Applied Multivariate Statistics.* New York, Wiley, 1990.

YOHANNES, Y., WEBB, P. *Classification and Regression Trees, CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity.* International Food Policy Research Institute (IFPRI), 1999. Disponível em: `http://www.ifpri.org/publication/classification-and-regression-trees-cart`, acesso em 20 jun. 2010.

ZHANG, H., SINGER, B. *Recursive Partitioning in the Health Sciences.* New York, Springer, 1999.

Apêndice A

Uso Combinado do Aprendizado Supervisionado e Não-supervisionado

Este Apêndice apresenta o uso combinado do aprendizado supervisionado e não-supervisionado através de uma aplicação, na área da Astronomia, das redes neurais do tipo SOM (Seção 2.5.2) e da análise de componentes principais (Seção 2.4.1), para o reconhecimento de padrões em dados de emissão de raios gama:

PEREIRA, B. DE B., RAO, C. R., OLIVEIRA, R. L., NASCIMENTO, E. M. Combining Unsupervised and Supervised Neural Networks Analysis in Gamma Ray Burst Patterns Classification, *Journal of Data Science*, v. 8, pp. 327-338, 2010.

Journal of Data Science **8**(2010), 327-338

Combining Unsupervised and Supervised Neural Networks in Cluster Analysis of Gamma-Ray Burst

Basilio de B. Pereira[1], Calyampudi R. Rao[2], Rubens L. Oliveira[1,3] and Emília M. do Nascimento[1]
[1]*Federal University of Rio de Janeiro,* [2]*Penn State University and AIMSCS and* [3]*Brazilian Naval Research Institute(IPqM)*

Abstract: The paper proposes the use of Kohonen's Self Organizing Map (SOM), and supervised neural networks to find clusters in samples of gamma-ray burst (GRB) using the measurements given in BATSE GRB. The extent of separation between clusters obtained by SOM was examined by cross validation procedure using supervised neural networks for classification. A method is proposed for variable selection to reduce the "curse of dimensionality". Six variables were chosen for cluster analysis. Additionally, principal components were computed using all the original variables and 6 components which accounted for a high percentage of variance was chosen for SOM analysis. All these methods indicate 4 or 5 clusters. Further analysis based on the average profiles of the GRB indicated a possible reduction in the number of clusters.

Key words: Bayesian regularization, clustering and classification, cross validation, multilayer perceptron, self organizing map, supervised and unsupervised networks.

1. Introduction

It is of great interest to astronomers to know whether the measurements on gamma-ray burst (GRB) can be characterized by a single probability distribution around some central value or as a mixture of probability distributions around different central values. Clustering is an exploratory data analysis (EDA) for investigating such problems by looking for groups of observed samples which are well separated using a suitable criterion. The ultimate aim is to seek for a physical interpretation of differences between the groups. An interesting example in a different context is the discovery of three clusters of the general population of individuals based on some blood tests for diabetes, one identified as diabetes free, and the other two representing individuals with 2 different types of diabetes A and

Table 1: Initial Variables

#	ID	Description	Mean of Log	S.d. of Log
1	T50	time measure representing the arrival of 50% of the flux	1.33	2.18
2	T90	time measure representing the arrival of 90% of the flux	2.29	2.19
3	F1	time-integrated fluence in spectral channels 20-50 keV	-15.61	2.11
4	F2	time-integrated fluence in spectral channels 50-100 keV	-15.15	1.99
5	F3	time-integrated fluence in spectral channels 100-300 keV	-13.94	1.82
6	F4	time-integrated fluence in spectral channels over 300 keV	-13.56	1.96
7	P64	peak flux measured in 64ms bins	0.74	1.02
8	P256	peak flux measured in 256ms bins	0.49	1.05
9	P1024	peak flux measured in 1024ms bind	0.05	1.14
10	T64	trigger threshold, i.e., number of counts in 64 ms required to trigger the second most brightly illuminated detector	4.20	0.05
11	T256	trigger threshold on the 256 ms timescale	4.89	0.07
12	T1024	trigger threshold on the 1024 ms timescale	5.59	0.09
13	Lat	galactic latitude	4.95	0.87
14	Lon	galactic longitude	4.95	0.94
15	FT	sum of the four fluencies (F1 + F2 + F3 + F4)	0.69	0.71
16	H32	spectral hardness, obtained from fluence relation F3/F2	1.21	0.64
17	H321	spectral hardness, obtained from relation F3/(F1+ F2)	-12.62	1.71

B (Reaven and Miller, 1979). Another example is the discovery of 2 clusters of individuals suggesting 2 types of cancers (Golub et al., 1999). Cluster analysis is a valuable tool in knowledge acquisition. In the literature there are two approaches to cluster analysis. One is parametric assuming a mixture of a given number of probability distributions such as multivariate normal. Another is nonparametric which offers a great flexibility in discovering the number of clusters and their shape without going through model selection procedures.

There are a number of methods of cluster analysis, a good review of which can be found in Jain, Murty and Flynn (1999) and Jiang *et al.* (2004). We use an unsupervised neural network known as SOM (Self Organizing Map) for finding

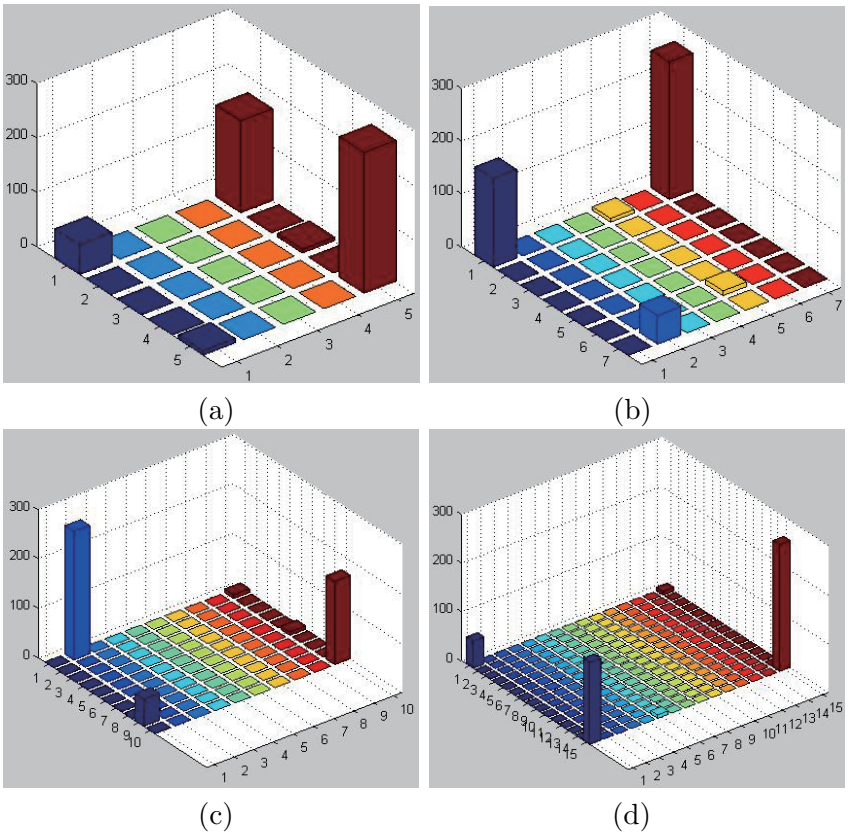


Figure 1: Clustering using Kohonen's maps of 5×5 nodes (a), 7×7 nodes (b), 10×10 nodes (c), and 15×15 nodes (d)

clusters and discuss methods of validating them by cross validation and profile analysis. We also propose two methods of reducing the number of variables for obtaining stable results. Some references to early work on cluster analysis of GRB are Mitrofanov *et al.* (1998), Bagoly *et al.* (1998), Mukherjee *et al.* (1998), Hakkila *et al.* (2000), and Rajaniemi and Mahonen (2002).

2. Cluster Analysis

2.1 Data

We consider the original BATSE 3B catalogue from the Compton Gamma Ray observatory, which is composed of 1122 GRB trigger samples with 14 measurements of astrophysical interest made on each sample. In addition we also list 3 other measurements usually considered in astrophysical research described in Murkerjee *et al.* (1998), Mitrofanov *et al.* (1998), Rajaneimi and Mahonen (2002). Since the computational complexity of the data mining process is not

increased dramatically by including additional variables, we used all 17 variables. The list of 17 variables is given in Table 1 with the mean values and standard deviations of log variables. Log transformation is made to reduce the variables to uniform scale.

2.2 Cluster analysis using SOM

There were 422 GRB samples with all variables present. A SOM was used for clustering these GRB patterns. Four different topologies were tried to test the clustering process. Figures 1a -1d show the number of patterns in each cluster over squared Kohonen's map of different dimensions with nodes: $25(5\times5), 49(7\times7), 100(10\times10)$ and $225(15\times15)$. As can be seen, the nodes representing the classes are well separated from each other in the 2 dimensional map provided by the topology. For a brief description of SOM and the underlying concepts, reference may be made to Rejaniemi and Mahonen (2000). All topologies clustered the 422 samples into 5 clusters designated as classes 1, 2, 3, 4 and 5. The fifth class had a small frequency and did not appear to be different from the fourth. They were combined to form one cluster as class 4.

2.3 Cross validation

Working with the 15×15 topology (the one which presented the maximum relative distance between classes), the input patterns were divided into two groups called *in sample set* (317 patterns) and an *out of sample* set (105 patterns) with a random algorithm using stratified sampling. A supervised MLP (Multilayer Perceptron) neural net with Bayesian regularization (see Mackay, 1992) was used to train the in sample set for classification of patterns into four classes. Ten different trainings were performed and the patterns in the out of samples were classified into four classes. The overall mean accuracy of classification was 92.4% and the error for each class is as given in Table 2.

Table 2: Misclassification error

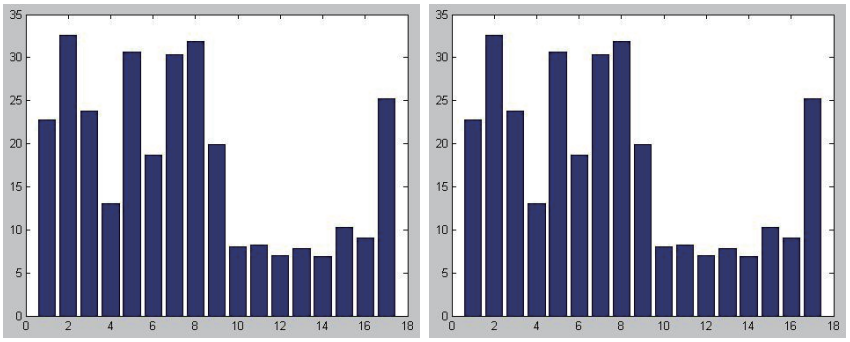
		Classified as class				Total	% Correct
		1	2	3	4		
Belonging	1	47	2	2	0	51	92.15
to class	2	2	14	0	0	16	87.50
	3	2	0	33	0	35	94.28
	4 + 5	0	0	0	3	3	100.00

It is seen that classes 1, 3 and 4 are well separated while 2 is not so well separated from 1. While this needs further discussion, we consider the four classes

to explain the method for reducing the number of variables.

2.4 Reduction of dimensionality

In multivariate analysis, one is faced with the curse of dimensionality as originally pointed out by Rao (1952) and referred to in the statistical literature as Rao's paradox. For obtaining stable results, a proper selection of variables has to be made. We suggest two procedures for this purpose, one of which is described in this section. The second is based on principal component analysis as detailed in the next section 3.



(a) (b)

Figure 2: Relative importance of each input variable: (a) before pruning (listed in table 1) and (b) after pruning ($T_{50}, T_{90}, F_1, F_2, F_3$, and F_4).

Figure 2 (a) presents, for each input variable of the feedforward neural network, the sum of the absolute values of the weights (S_i) connecting the corresponding input to the hidden layer neurons. Taking the mean (M) and the standard deviation (SD) of theses sums and using as threshold (T) the value $T = M - SD$, we eliminated the variables whose S_i were below T. The neural network was trained 10 times, for randomly chosen sets of the initial weights, and the pruning criterion was used to confirm the eliminated variables. The average of the misclassification error for these 10 samples will be denoted by AV_j.

After eliminating variables, further 10 training samples were used and the misclassification errors were computed. If the average of these errors (AV_{j+1}) was more than the AV_j value, then the variables would be definitively abandoned.The procedure is repeated iteratively until the elimination of variables does not improve the misclassification error.

Figure 2 (b) shows the relative importance of each of the remaining input variables that were considered most relevant for the classification process (respectively $T_{50}, T_{90}, F_1, F_2, F_3$ and F_4). For this final configuration, the misclas-

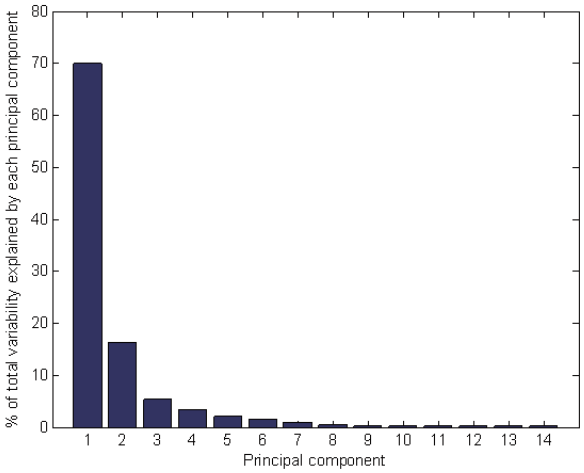


Figure 3: Percent of the total variability explained by each principal component.

sification error was 5.9% for the *out of sample* set and 1.4% for the *in sample* set).

Considering six variables, the number of the available observations (without missing values) increased from 422 to 632. Using again a 15×15 topology for SOM, now for the six remaining input variables and for 632 patterns, the classes and frequencies found were similar to the classes obtained using all seventeen input variables.

A feedforward neural network trained with the final six variables and 498 in sample and 134 out of sample observations resulted in an out of sample misclassification error of 5.9% compared with 7.6% with the 17 variables of the initial network with 317 in sample and 105 out of sample observations.

However, considering that the objective is to compare this methodology with the next one (described in section 3), only the 422 patterns initially considered will be used to compare the two methods.

3. Principal Component Analysis

A second approach to reduction of dimensionality is PCA (Principal Component Analysis) where the variables are replaced by a smaller number of linear functions of the variables. In computing the principal components only the first 14 variables of Table 1 are used. It may be noted that the last 3 variables of Table 1 are functions of the variables F_1, F_2, F_3, and F_4 in the list. The computations made on 422 samples where all the variables are available provided 14

linear combinations of the variables with the associated eigen values as indicated in Figure 3. The first six principal components accounted for 98% of the total variance, and the SOM was used for clustering based on these components only. The analysis provided the same 5 classes as discussed in Section 2.

Table 3: Principal components according to PCA algorithm

Variables	Components					
	1st	2nd	3rd	4rd	5rd	6rd
log T50	0.414	-0.449	-0.103	-0.021	-0.033	-0.363
log T90	0.432	-0.388	-0.074	-0.049	-0.007	-0.355
log F1	0.441	0.016	0.288	0.079	0.015	0.513
log F2	0.420	0.068	0.253	0.063	-0.009	0.337
log F3	0.374	0.197	0.029	-0.001	-0.007	0.039
log F4	0.299	0.432	-0.823	-0.005	0.015	0.086
log P64	0.060	0.405	0.210	-0.049	-0.035	-0.338
log P256	0.097	0.387	0.231	-0.052	-0.028	-0.370
log P1024	0.169	0.321	0.246	-0.053	0.009	-0.290
log T64	0.000	-0.001	-0.001	-0.004	0.001	0.004
log T256	0.000	0.000	-0.001	-0.007	-0.001	0.003
log T1024	0.001	0.002	-0.001	-0.013	0.002	-0.007
log Lon	0.020	-0.015	0.022	-0.575	0.815	0.048
log Lat	-0.002	0.021	-0.001	0.805	0.577	-0.137
% variance	70.003	16.267	5.150	3.448	1.983	1.569
% cumulative	70.003	86.270	91.421	94.869	96.852	98.421

From the results of Table 3, we conclude that variables $T_{50}, T_{90}, F_1, F_2, F_3$ and F_4 are the most important. This result agrees with the previous one obtained using the MLP with a regularization technique which showed these same six variables as the most relevant to the classification process.

Using this method, we obtained the same 5 classes of the previous analysis, with the same patterns in each class. The full table with the composition of each class is available from the authors. Labeling the classes found by SOM's training process with numbers 1 to 5, it is possible to draw the patterns into graphs with the first versus the second and the third principal components provided by PCA analysis. These classes are clearly seen in Figure 4 and 5, where there is evidence of three classes (1,2 and 3). The status of classes 4 and 5 is not clear. However, some possibilities are that they may be considered as separate classes, class 5 may be merged with class 3, and 4 with 2. The profile analysis carried out in the next section also suggests similar grouping.

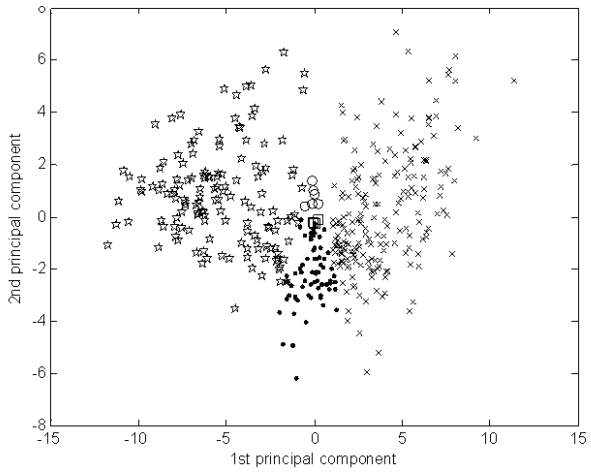


Figure 4: Patterns classified by SOM projected in their 2 main principal components (supplied by PCA) – class 1 represented by 'x', class 2 by '•', class 3 by star, class 4 by 'o' and class 5 by '□'.

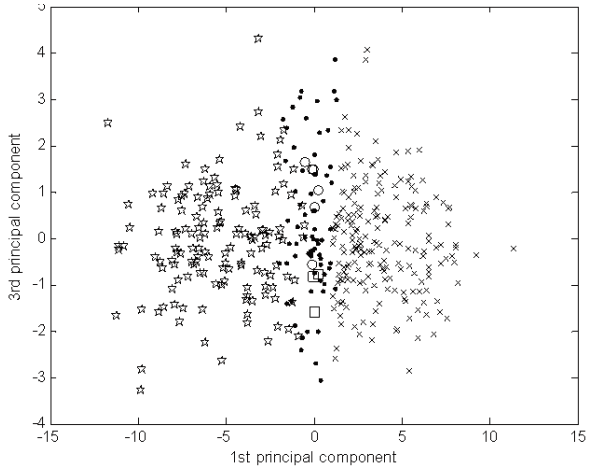


Figure 5: Patterns classified by SOM projected in their first and third principal components – class 1 represented by 'x', class 2 by '•', class 3 by star, class 4 by 'o' and class 5 by '□'.

4. Graphical Evaluation of the Classes

There is no recommended statistical method as the best for evaluating the validity and the number of clusters determined by using one or more of the numerous algorithms available for cluster analysis. See Sugar and James (2003) and Jiang *et al.* (2004). Figure 6 gives the distortion curves recommended in

Sugar and James (2003), which suggests about 4 classes. Another suggested method is to examine the profiles of the patterns in different classes, which in statistical literature is also known as the plot of parallel coordinates of individuals and mean values as shown in Figure 7. It is seen that Classes 1 and 3 are distinct with class 2 occupying an intermediate position. The positions of classes 4 and 5 are not clear. It is interesting to see that the four classes differ mainly in mean values of the six variables chosen for clustering in Section 3.

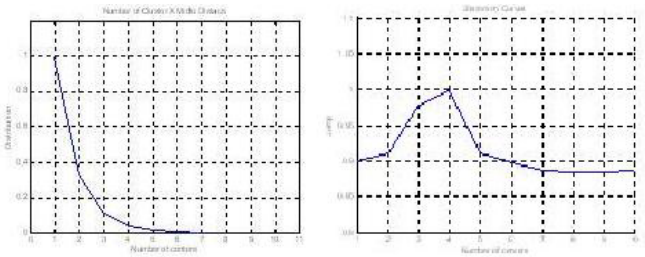
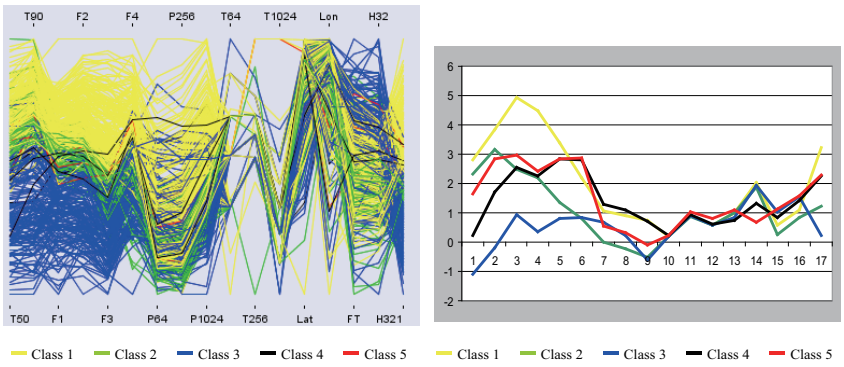


Figure 6: Number of cluster x middle distance and distortion curve



(a) (b)

Figure 7: Profiles and means. Panel (a): Parallel coordinate diagram, Panel (b): Means of variables (centralized for comparisons).

5. Conclusion

Our study indicates the following:

a) The profile plot and the scatter plot of the first two principal components indicate a clear separation between Classes 1 and 3. Patterns in Class 1 are characterized by long duration, bright fluency and soft spectrum while Class 3, by short duration, faint fluency and hard spectrum.

There is some overlap between Classes 1 and 2 in the profile plot, but the distinctiveness of Class 2 is brought out in the plot of principal components. Patterns in this class are characterized by intermediate duration and fluence, and hard spectrum. The positions of Classes 4 and 5 are not clear. However, the profile plots of classes 4 and 5 appear to be similar. Patterns in these classes can be characterized by intermediate duration, fluency and spectrum.

b) The means of the variables T50, T90, F1, F2, F3, F4 and H321 of the 3 clusters are well differentiated while the means of the other 10 variables P64, P256, P1024, T64, T256, T1024, Lat, Lon, Ft and H32 are not. The latter variables may not be useful in predicting the class to which a future GRB belongs. Any physical interpretation of the clusters should take this into account.

d) SOM seems to be an appropriate tool for clustering and graphical display of the results.

e) The choice of the dominant principal components is a computationally convenient way of reducing the curse of dimensionality due to a large number of variables in cluster analysis and classification problem

f) SOM provides non overlapping clusters and the distinction between the Classes 1 and 2 cannot be easily specified. A parametric approach such as fitting a mixture model may reveal three components as demonstrated in the paper by Mukherjee *et al.* (1998).

References

Bagoly, Z. *et al.* (1998). A principal component analysis of the 3B gamma-ray burst data. *The Astrophysical Journal* **498**, 342-348.

Golub, T. R. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Hakkila, J. *et al.* (2000). Gamma-ray class properties. *The Astrophysical Journal* **538**, 165-180.

Hoffman, P., Grinstein, G. and Pinkney, D. (1999). Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *Proceedings of the 1999 Workshop on New Paradigms in Information Visualization and*

Manipulation in Conjunction with the Eighth ACM Internation Conference on Information and Knowledge Management (Kansas City, Missouri, United States, November 02 - 06, 1999). NPIVM '99. ACM, New York, NY, 9-16. DOI= http://doi.acm.org/10.1145/331770.331775

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Survey* **3**, 264-323.

Jiang, D., Tang, C. and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Transaction on Knowledge Discovery and Data Engineering* **16**, 1370-1386.

MacKay, David J. C. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation* **4**, 448-472.

Mitrofanov, I. *et al.* (1998). Generic signatures of the time profiles of BATSE cosmic gamma-ray bursts. *The Astrophysical Journal* **504**, 925-934.

Mukherjee, S. *et al.* (1998). Three types of gamma ray bursts. *The Astrophysical Journal* **508**, 314-327.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. Wiley

Rajaneimi, H. and Mahonen, P. (2002). Classifying gamma-ray burst using self-organizing maps. *The Astrophysical Journal* **566**, 202-209.

Reaven, G.M. and Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using multidimensional analysis. *Diabetologia* **16**, 17-24.

Smyth, P. (1996). Clustering using Mont Carlo cross-validation. *Proc 2nd International Conference on Knowledge Discovery and Data Mining AIII Press*, 126-133.

Sugar, A. and James, M. (2003). Finding the number of cluster in a data set :an information theorectical approach. *Journal of the American Statistical Association* **98**, 750-763.

Received April 30, 2009; accepted August 26, 2009.

Basilio de B. Pereira
Federal University of Rio de Janeiro
Pos Graduate School of Engineering (COPPE)
Faculty of Medicine and University Hospital Clementino Fraga Filho
COPPE/UFRJ, CP 68507 , CEP 21941-972, Rio de Janeiro, RJ, Brazil
basilio@hucff.ufrj.br

Calyampudi R. Rao
Penn State University and AIMSMS
326 Thomas Building, University Park, PA 16802,USA
crr1@psu.edu

Rubens L. Oliveira
Brazilian Naval Research Institute(IPqM) and
Federal University of Rio de Janeiro
Pos Graduate School of Engineering (COPPE)
COPPE/UFRJ
Caixa Postal 68506 - CEP: 21945-970 ,Rio de Janeiro - RJ, Brazil
rubenliv@yahoo.com

Emília M. do Nascimento
Federal University of Rio de Janeiro
Pos Graduate School of Engineering (COPPE)
COPPE/UFRJ
CP 68507 , CEP 21941-972, Rio de Janeiro, RJ, Brazil
emilia@pep.ufrj.br

Apêndice B

Árvores de Sobrevida

O artigo, apresentado neste Apêndice, mostra uma aplicação das árvores de sobrevida (Seção 2.3.3) e do modelo de riscos proporcionais de Cox (Seção 2.2), em dados de pacientes na fila de transplante hepático:

NASCIMENTO, E. M., PEREIRA, B. DE B., BASTO, S. T., RIBEIRO FILHO, J. Survival Tree and Meld to Predict Long Term Survival in Liver Transplantation Waiting List, *Journal of Medical Systems*, 2010. DOI: 10.1007/s10916-010-9447-6.

J Med Syst
DOI 10.1007/s10916-010-9447-6

ORIGINAL PAPER

Survival Tree and Meld to Predict Long Term Survival in Liver Transplantation Waiting List

Emília Matos do Nascimento · Basilio de Bragança Pereira · Samanta Teixeira Basto · Joaquim Ribeiro Filho

Received: 29 September 2009 / Accepted: 8 February 2010

Abstract MELD score is a formula based on laboratory variables used as a predictor of short-term mortality index in cirrhotic patients. It is applied to allocate patients in liver transplantation waiting list in many countries. However, MELD score cutoff point accuracy to predict long term mortality has not been statistically evaluated. The aim of this study was to analyze the MELD score and other variables related to long-term mortality using a new model: the Survival Tree analysis. The variables considered in this study were obtained at the time of liver transplantation list enrollment. The graphical representation of the survival trees showed that MELD 16 was the most statistically significant mortality cutoff point. The results were compatible with the MELD cutoff point reported in the clinical literature. This methodology can be extended to identify significant cutoff points related to other diseases whose severity is not necessarily expressed by MELD.

Keywords Decision trees · Health status indicators · Liver transplantation · Survival analysis · Waiting lists

E. M. do Nascimento (✉)
Federal University of Rio de Janeiro, COPPE - Postgraduate School of Engineering; HUCFF - University Hospital Clementino Fraga Filho,
Rio de Janeiro, Brazil
e-mail: emilia@pep.ufrj.br

B. B. Pereira
UFRJ - Federal University of Rio de Janeiro,
HUCFF - University Hospital Clementino Fraga Filho; FM - School of Medicine; COPPE - Postgraduate School of Engineering,
Rio de Janeiro, Brazil

S. T. Basto · J. Ribeiro Filho
UFRJ - Federal University of Rio de Janeiro,
HUCFF - University Hospital Clementino Fraga Filho;
FM - School of Medicine,
Rio de Janeiro, Brazil

Introduction

The Model for End-Stage Liver Disease (MELD) [1] score was described as a short term mortality index used to predict three month mortality in patients who underwent transjugular intrahepatic portosystemic shunt (TIPS) insertion, ranging from 0 to 40. It was subsequently applied to allocate liver grafts in liver transplantation list in the United States and several countries, since February 2002 [2]. Many countries use subjective local criteria or UNOS based policy to allocate liver grafts according to liver disease severity [3].

The liver transplantation waiting list time varies significantly among various centers but usually reflects a gap between the donor liver pool and the demand for transplant [4]. The longer waiting time results in a higher mortality rate [5]. There is a worldwide tendency to increase waiting list time, due to organ shortage.

There are several factors related to liver transplantation waiting list mortality reported in the literature as age, gender, blood type and disease etiology [6]. The MELD advantage to allocate organs is its reproducibility and absence of subjective criteria as ascitis or encephalopathy.

Many authors have described MELD as an independent tool related to short term mortality in the transplantation waiting list and tried to determine a threshold to assess prognosis and mortality in this setting [7, 8]. However, MELD score accuracy to predict long term mortality has not been statistically evaluated.

Considering that waiting list times often exceeds one year in many countries, the motivation of this study is to analyze the performance of the variables used to allocate liver organs such as MELD score as much as other clinical significant variables to predict mortality in a long term waiting list for cirrhotic patients. To attain this goal, a new useful statistical method, the Survival Tree analysis, is

Published online: 03 March 2010

proposed to establish a cutoff point of MELD, as well as other variables, that better predicts this long term mortality. The interactions between the explanatory variables are also analyzed.

Materials and methods

Data set

All patients referred for liver transplantation waiting list, in a single center, during a 10 year period were evaluated for inclusion in the study. Due to available data, a total of 529 patients were considered in this study.

Data were obtained from the patient inclusion registration form and from the hospital's internal system of patient registration and organized in excel for posterior analysis.

The variables obtained at the time of liver transplantation list enrollment and considered in this study were: gender, age, blood type, body mass index, liver disease etiology, hepatocellular carcinoma, waiting time for transplant (in days) and MELD. The MELD score formula [1] were:

$$MELD = 3.8\log(b) + 11.2\log(INR) + 9.6\log(cr) + 6.4*et$$

where b is total serum bilirubin [mg/dL]; INR is the International Normalized Ratio for prothrombin time; cr is the serum creatinine [mg/dL]; and et is the etiology of cirrhosis (0 if cholestatic or alcoholic, 1 otherwise).

Mortality on the waiting list was the outcome. Dropout list, transplantation or still waiting in the transplantation list at the end of the study were considered as censored data.

The statistical approach used was the Survival Tree developed by Hothorn et al. [9] and the Proportional Hazard Model of Cox [10]. The implementation was done using R [11] packages.

Recursive partitioning

A learning set L consists of m covariates $X = (X_1,..., X_m)$ of a sample space χ and a response Y of a sample space Y. Let it be a learning set L used to form a predictor $\varphi(x, L)$, i.e., if the input is x the answer y will be predicted by $\varphi(x, L)$.

So, the conditional distribution $D(Y|X)$ of the response given covariates X depends on a function f of the covariates $D(Y|X) = D(Y|X_1,\ldots,X_m) = D(Y|f(X_1,\ldots,X_m))$, with the restriction that the partition is based on the regression relationships so that the covariate space χ is partitioned in r disjoint cells $B_1,..., B_r$.

The regression model will be fitted based on a learning sample L_n composed of n independent and identically distributed observations.

The association between the response Y and covariates X_j, $j = 1, \ldots, m$, is measured by the following linear statistics:

$$T_j(L_n, w) = vec\left(\sum_{i=1}^{n} w_i g_j(X_{ji}) h(Y_i, (Y_1, ..., Y_n))^T\right) \in \Re^{p_j q}$$

where $g_j : X_j \rightarrow \Re^{p_j}$ is a non-random transformation of the covariate X_j ; $h : Y \times Y^n \rightarrow \Re^q$ is the influence function that depends on the responses $(Y_1, ...,Y_n)$ in a permutation symmetric way. For survival data the influence function may be the logrank taking censoring into account. A weighted Kaplan-Meier curve for the weights $w(X)$ can serve as prediction; and vec is the operator that converts a p_j x q matrix into a p_jq column vector by column-wise combination.

The distribution of $T_j(L_n, w)$ under the partial hypotheses H_j^0 depends on the joint distribution of Y and X_j , which is unknown.

Most algorithms for the construction of classification or regression trees algorithm follow a general rule [12]: first, partition the observations by univariate splits in a recursive way; and second, fit a constant model in each cell of the resulting partition.

The conditional inference trees [9] embed recursive binary partitioning into the well defined theory of permutation tests developed by Strasser and Weber [13], following three steps. First, considering the case weights w, test the global null hypothesis of independence between any covariates and the outcome. If this hypothesis cannot be rejected, then stop. Otherwise select the covariate with strongest association to the outcome, which measure is related to a p-value to test the partial null hypothesis of a covariate and the outcome. Second, choose a set of the predictor to implement binary split in the selected covariates defined by the left and right case weights (w_{left} and w_{right}). Third, the two previous steps are repeated recursively until the stop criterion is achieved.

The algorithm stops if the global null hypothesis of independence between the response Y and any of the m covariates cannot be reject at a pre-specified nominal level α, which in our case was 0.05. Otherwise the association between the response and each of the m covariates is measured by test statistics or p-values that indicate the deviations from the partial hypotheses H_j^0.

Proportional hazard model

A Proportional Hazard Model of Cox [10] will also be employed to investigate the association between survival time and the explanatory variables. The model assumes that for an individual with vector x of covariates, the hazard rate at time t is given by

$$h(t;x) = h_0(t)\exp\left(\beta_x'\right)$$

where $h_0(t)$ is the underlying hazard function at time t for $x = 0$ (i.e, all covariates at their appropriate reference level) and β is a vector of unknown coefficients of covariates effects. In the present case the dependent variable t is the time measured in days since the entrance of each patient in the transplant queue until death or censoring. The explanatory variables were: gender, age, blood type, body mass index, liver disease etiology, hepatocellular carcinoma (HCC), and MELD.

Results

From the 529 patients in the data base, 61% were male. The mean age was 51±13 years old. The most frequent etiology for liver disease was chronic hepatitis C (47%), alcoholic liver disease (17%) and cryptogenic cirrhosis (10%). Regarding general outcome, 36% died, and 64% were censored, from which 8% left the transplant list, 14% had been submitted to a liver transplant, and 42% are still in list. The median follow-up in liver transplantation waiting list was 1011 days.

The application of the nonparametric regression tree to the data is presented in Figs. 1, 2, 3 and 4. A graphical representation of the survival tree for the 529 patients in the liver transplantation waiting list was implemented using the party add-on package [9] to the R system [11] for statistical computing.

The main cutoff point in each tree corresponds to the p-value related to that variable in the Cox regression and p-value corresponds to the log-rank test [12].

In Fig. 1 one can observe the MELD cutoff at 16. This survival tree also presents some other cutoff statistically significant. Figure 1 also shows that the higher the MELD score, the higher the mortality.

Figure 2 shows that the first important cutoff is related to MELD at 16 and also presents the interaction with age where the cutoff corresponds to 33.2 years.

Figure 3 shows again the main cutoff corresponding to MELD at 16 and also the relevant interaction with hepatocellular carcinoma diagnosis (HCC). It is worth noting, in node 5, a group of 13 patients where one can observe high mortality in spite of meld between 12 and 16.

Finally, Fig. 4 presents a decision cutoff with three variables: MELD (cutoff at 16), age, and HCC diagnosis. Figure 4 also shows three groups of interesting features: the first one with 13 patients (node 5) where one can observe high mortality, meld between 12 and 16, and HCC diagnosis; the second group with 22 patients (node 8) with low mortality, meld greater than 16, and age less than 33 years old; and the third with 126 patients (node 9) with

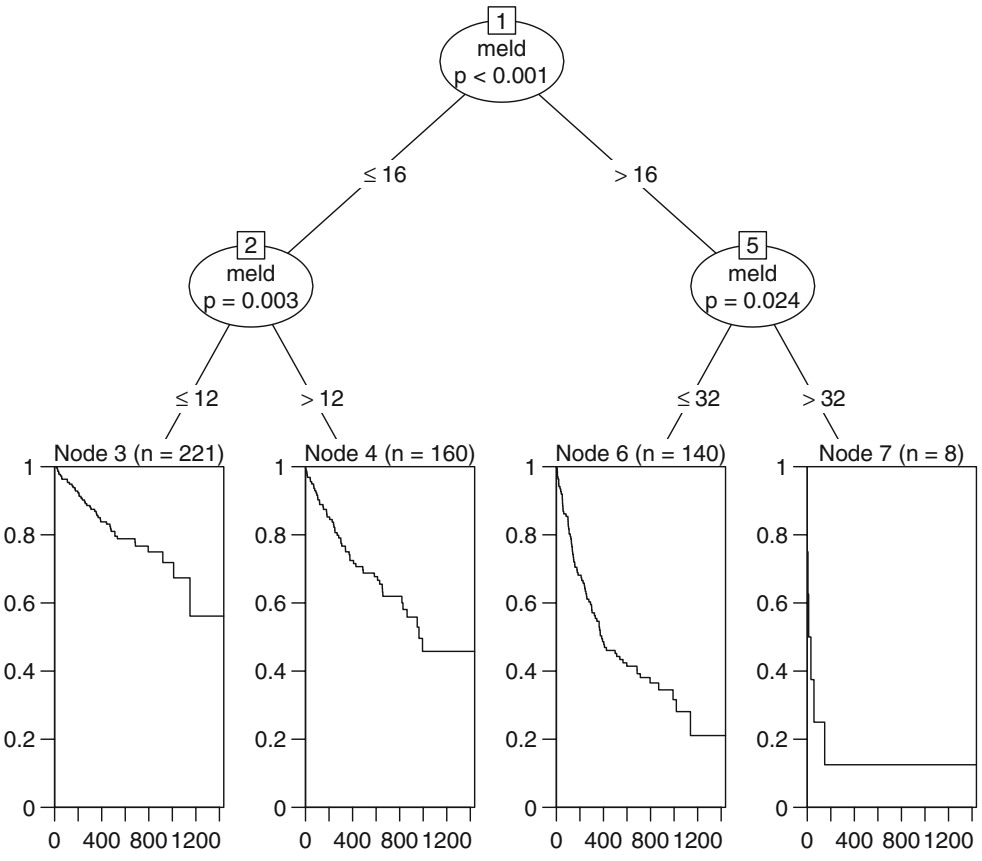


Fig. 1 Survival Tree (MELD)

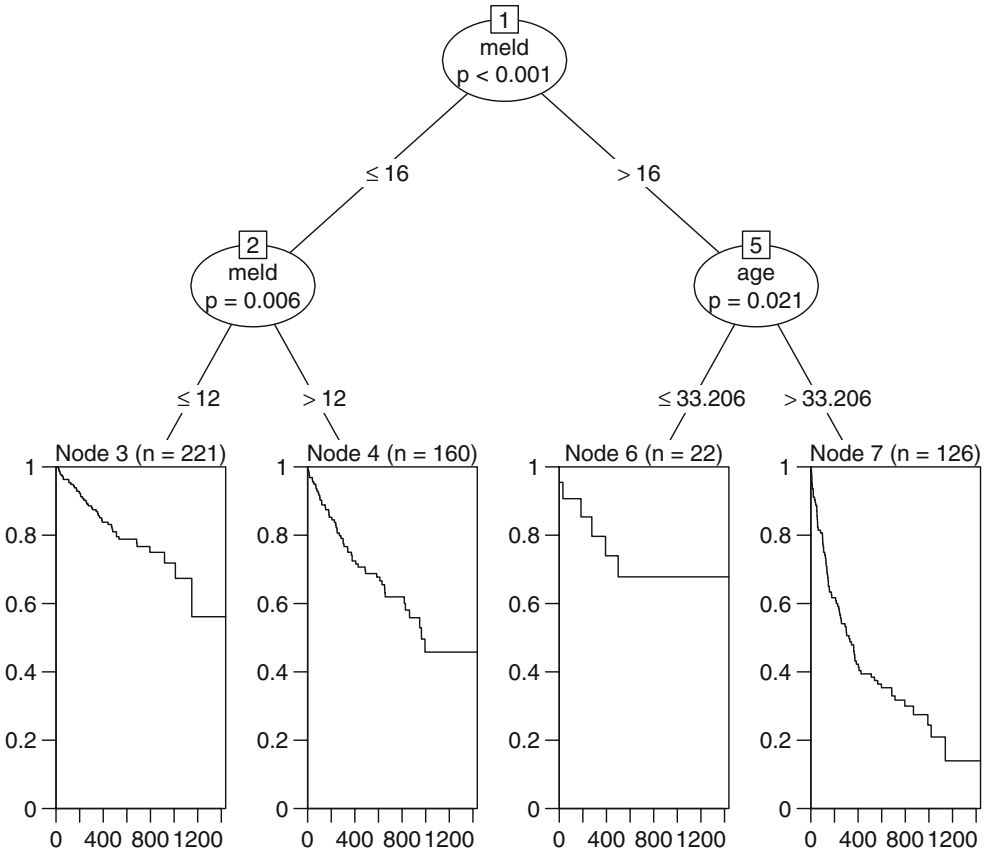


Fig. 2 Survival Tree (Interaction between MELD and age)

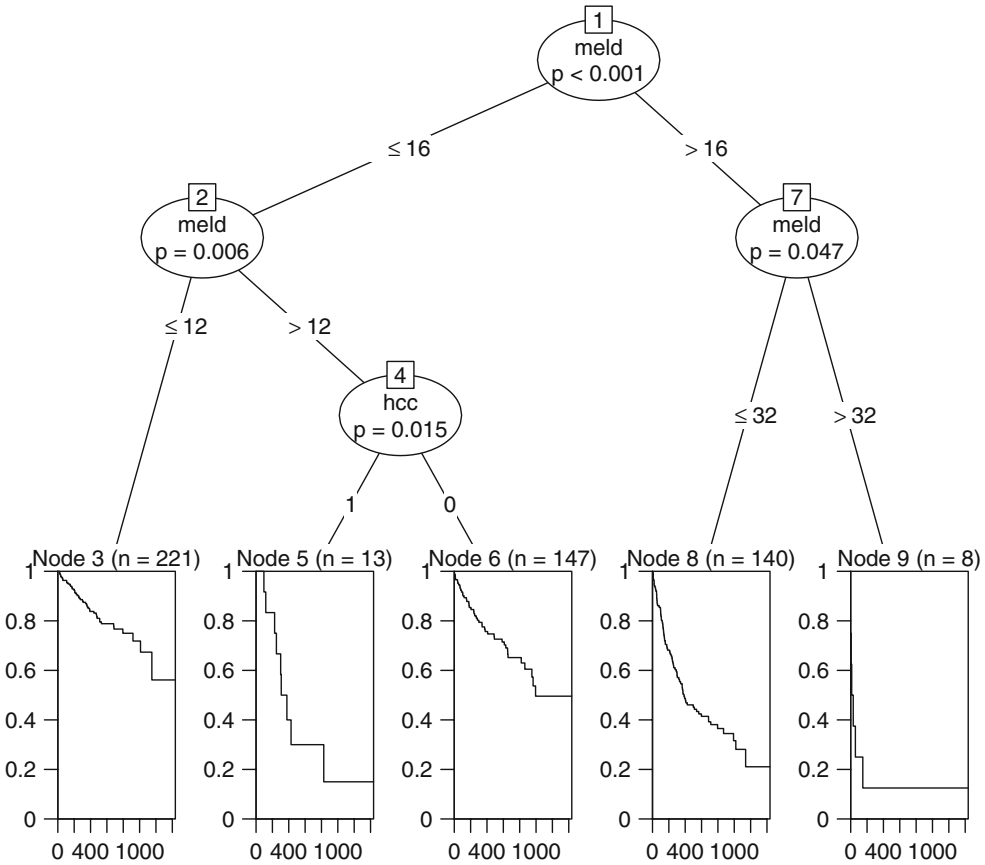


Fig. 3 Survival Tree (Interaction between MELD and HCC)

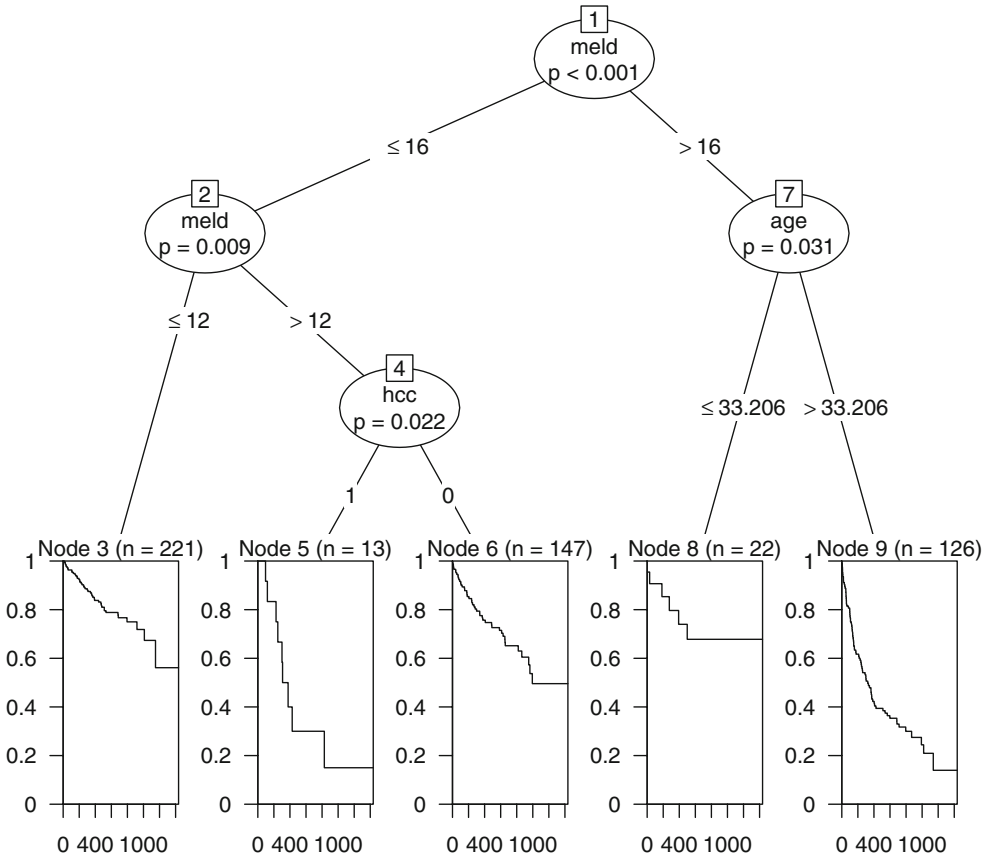


Fig. 4 Survival Tree (Interaction between MELD, age and HCC)

high mortality, MELD above 16, and age older than 33 years.

In addition, Table 1 presents the results of Cox Proportional Hazard model which also show the significance of these three variables, also indicating that the survival tree shown in Fig. 4 is the best model.

The other variables in the data set did not show any interaction with MELD, when the Cox model was used.

Discussion

The aim of an organ allocation policy is to promote maximum survival not only in liver transplantation programs by reducing waiting time, death and dropout list rates, but also reducing overall mortality for patients with end-stage organ failure.

Since the introduction of MELD score for organ allocation in liver transplantation in 2002, the cutting point of 15 has been described to allow patients onto the waiting list [14].

No system is perfect and requires constant modification. The MELD score still has limitations for adequate evaluation in around 20% of patients listed with hepatocellular carcinoma, cholestatic or metabolic diseases and cirrhosis decompensated by encephalopathy, ascites or digestive hemorrhage [15].

Boin et al. [16] reported Child-Pugh score as a mortality index in long term waiting list, showing no difference from MELD score.

The current study describes demographical and clinical characteristics of a liver transplantation program in a developing country, with long waiting time for transplantation. That setting could reflect more accurately the natural history of advanced liver disease, i.e., without intervention.

In this paper MELD score cutoff to predict long term mortality in liver transplantation waiting list was statistically evaluated. Survival Tree analysis and MELD was used to predict long term mortality.

Our statistical results reinforce a similar mortality MELD cutoff point reported for short term basis in the medical literature [1]. This information corroborates the MELD allocation system adoption in regions with longer waiting times for liver transplantation, as observed in our country. The use of the Survival Tree analysis along with the

Table 1 Cox model for variables MELD, age and HCC

	coef	se(coef)	z	p
age	0.0208	0.00645	3.23	0.0013
meld	0.1097	0.01094	10.03	0.0000
hcc1	0.5259	0.21502	2.45	0.0145

adoption of the MELD criterion can provide a more fair policy for allocation of liver grafts.

In clinical research, statistical modeling is a powerful tool to investigate the relationship between an outcome and a set of covariates. A Proportional Hazard Model of Cox [10] is used in study of prognostic factors to assess the association between survival times and the covariates effects. The Cox model does not require any assumptions about the probability function of the data. On the other hand, the Survival Trees [9] are useful to find the cutoff point on covariates and have the additional advantage of attractiveness due to their easy visualization, intuitive interpretation of the results, and immediate correlation to individual clinical assessment.

Since Cox regression model indicates MELD, age and HCC as important covariates and the survival tree shown in Fig. 4 pointed out the same covariates, we can choose this as the most representative survival tree for predicting long term mortality based on our data set.

Besides MELD mortality cutoff at 16, obtained in this study, additional decision criterion is a further information obtained from the survival tree, pointing other high risk sub groups in our casuistic: The patients with MELD between 12 and 16 and HCC diagnosis, and also MELD above 16 and age older than 33 years.

Survival Trees can also provide additional information for diseases whose severity are not directly related to the MELD value, such as hepatocellular carcinoma, cholestatic or metabolic diseases and cirrhosis decompensated by encephalopathy, ascites or digestive hemorrhage.

References

1. Kamath, P. S., Wiesner, R. H., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., et al., A model to predict survival in patients with end-stage liver disease. *Hepatology* 33 (2):464–470, 2001.
2. Edwards, E. B., and Harper, A. M., Application of a continuous disease severity score to the OPTN liver waiting list. Clin. Transpl. 19–24, 2001.
3. Fink, M. A., Angus, P. W., Gow, P. J., Berry, S. R., Wang, B. Z., Muralidharan, V., et al., Liver transplant recipient selection: MELD vs. clinical judgment. *Liver. Transpl.* 11(6):621–626, 2005.
4. Zapata, R., Innocenti, F., Sanhueza, E., Humeres, R., Rios, H., Suarez, L., et al., Clinical characterization and survival of adult patients awaiting liver transplantation in Chile. *Transplant. Proc.* 36(6):1669–1670, 2004.
5. Ransford, R., Gunson, B., Mayer, D., Neuberger, J., and Christensen, E., Effect on outcome of the lengthening waiting list for liver transplantation. *Gut* 47(3):441–443, 2000.
6. Fink, M. A., Berry, S. R., Gow, P. J., Angus, P. W., Wang, B.-Z., Muralidharan, V., et al., Risk factors for liver transplantation waiting list mortality. *J. Gastroenterol. Hepatol.* 22(1):119–124, 2007.
7. Adler, M., DeGendt, E., Vereerstraeten, P., Degré, D., Bourgeois, N., Boon, N., et al., Value of the MELD score for the assessment of pre- and post-liver transplantation survival. *Transplant. Proc.* 37(6):2863–2864, 2005.
8. Lee, Y. M., Fernandes, M., DaCosta, M., Lee, K. H., Sutedja, D., Tai, B. C., et al., The MELD score may help to determine optimum time for liver transplantation. *Transplant. Proc.* 36 (10):3057–3059, 2004.
9. Hothorn, T., Hornik, K., and Zeileis, A., Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15(3):651–674, 2006.
10. Cox, D. R., Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society Series B* 34(2):187–220, 1972.
11. R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2009.
12. Everitt, B. S., and Hothorn, T., *A Handbook of Statistical Analyses Using R*. Chapman & Hall-CRC Press, Boca Raton, 2006.
13. Strasser, H., and Weber, C., On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics* 8:220–250, 1999.
14. Perkins, J. D., Halldorson, J. B., Bakthavatsalam, R., Fix, O. K., Carithers, R. L., Jr., and Reyes, J. D., Should liver transplantation in patients with model for end-stage liver disease scores ≤14 be avoided? A decision analysis approach. *Liver. Transpl.* 15(2):242–254, 2009.
15. Thuluvath, P. J., Maheshwari, A., Thuluvath, N. P., Nguyen, G. C., and Segev, D. L., Survival after liver transplantation for hepatocellular carcinoma in the model for end-stage liver disease and pre-model for end-stage liver disease eras and the independent impact of hepatitis C virus. *Liver Transpl.* 15(7):754–762, 2009.
16. Boin, I. F. S. F., Leonardi, M. I., Pinto, A. O., Leme, R. S. R., Udo, E., Stucchi, R. S. B., et al., Liver transplant recipients mortality on the waiting list: long-term comparison to Child-Pugh classification and MELD. *Transplant. Proc.* 36(4):920–922, 2004.

Apêndice C

Curvas de Sobrevida

As curvas de sobrevida (Seção 2.2.1) foram utilizadas para avaliar a eficácia do tratamento da síndrome do olho seco em pacientes infectados pelo HIV com suplementação lacrimal ou com suplementação lacrimal associada à terapia anti-inflamatória. Para comparação das curvas de sobrevida foram utilizados os testes Log-rank, Tarone-Ware e Peto-Prentice (Seção 2.2.3):

BARRETO, R. P. P., BIANCARDI, A. L., NASCIMENTO, E. M., PEREIRA, B. DE B., MORAES JR, H. V. Uso de Ciclosporina 0,05% Tópica no Tratamento do Olho Seco de Pacientes Portadores do Vírus HIV. *Revista Brasileira de Oftalmologia*, v. 68, pp. 83-89, 2009.

Uso de ciclosporina 0,05% tópica no tratamento do olho seco de pacientes portadores do vírus HIV

Topical cyclosporine 0.05% for the treatment of dry eye disease in patients infected with the human immunodeficiency virus

Rodrigo de Pinho Paes Barreto[1], Ana Luiza Biancardi[2], Emília Matos Nascimento[3], Basílio de Bragança Pereira[4], Haroldo Vieira de Moraes Jr[5]

RESUMO

Objetivo: O presente estudo comparou a eficácia do tratamento da síndrome do olho seco em pacientes infectados pelo HIV com suplementação lacrimal (carboximetilcelulose sódica 0,5%) ou com suplementação lacrimal associada à terapia anti-inflamatória (carboximetilcelulose sódica 0,5% e ciclosporina 0,05%). **Métodos:** Vinte pacientes portadores do vírus HIV provenientes do ambulatório de Infectologia do Hospital Universitário Clementino Fraga Filho da Universidade Federal do Rio de Janeiro foram selecionados. O diagnóstico de síndrome do olho seco foi baseado no questionário para olho seco *(Ocular Surface Disease Index* - OSDI®), teste de Schirmer I, tempo de ruptura do filme lacrimal e coloração da superfície ocular com rosa bengala a 1%. Os pacientes foram distribuídos em dois grupos com dez pacientes (20 olhos) e acompanhados durante seis meses. O grupo I foi tratado com colírio de carboximetilcelulose sódica 0,5% e o grupo II foi tratado com a associação de carboximetilcelulose sódica 0,5% e ciclosporina 0,05% tópica. **Resultados:** Comparando os resultados no início e no final do tratamento, a associação com ciclosporina 0,05% mostrou-se superior ($p < 0,05$) para o teste de Schirmer I. **Conclusão:** O presente estudo sugere que o uso de ciclosporina 0,05% tópica melhora a produção lacrimal em pacientes infectados pelo HIV, apresentando um importante papel como adjuvante no tratamento do olho seco nestes pacientes.
ClinicalTrials.gov Identifier: NCT00797030

Descritores: Síndrome do olho seco/quimioterapia; Ceratoconjuntivite seca/quimioterapia, Ciclosporina/uso terapêutico; Lágrimas artificiais/administração & dosagem; Administração tópica; Soropositividade para HIV

[1]Mestre em Oftalmologia pela Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro (RJ), Brasil;
[2]Mestre em Oftalmologia pela Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro (RJ), Brasil;
[3]Pós-Graduanda nível Doutorado em Bioestatística da Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro (RJ), Brasil;
[4]Professor Titular do Departamento de Bioestatística da Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro (RJ), Brasil;
[5]Professor Associado do Departamento de Oftalmologia da Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro (RJ), Brasil.

Trabalho realizado no Hospital Universitário Clementino Fraga Filho – Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro (RJ), Brasil.

Recebido para publicação em: 27/11/2008 - Aceito para publicação em 25/4/2009

INTRODUÇÃO

A síndrome do olho seco se refere a uma doença da superfície ocular com diversas etiologias que frequentemente coexistem. A prevalência precisa da síndrome do olho seco é difícil pela falta de critérios diagnósticos concisos e pela subjetividade dos sintomas [1]. Estudos prévios estimaram prevalências compreendidas entre 5,5% e 33,7% [2-10].

Em 1995, encontros realizados no *National Eye Institute (NEI)* propuseram que o olho seco pode ser decorrente de uma deficiência da produção aquosa lacrimal ou de uma evaporação excessiva, com dano à superfície ocular interpalpebral e sintomas oculares [11].

Em 2003, a classificação tripla de Madri propôs que o diagnóstico do olho seco fosse baseado em três parâmetros: etiologia, histopatologia e gravidade clínica [12].

Em 2006, o painel *Delphi* propôs um consenso para classificação e tratamento do olho seco. Um novo termo para a síndrome do olho seco foi sugerido: síndrome de disfunção lacrimal. [13]

Em 2007, foram publicados os resultados do *International Dry Eye Workshop (DEWS)* [14], com o objetivo de ampliar os conceitos estabelecidos pelo *NEI* (1995) [11]. O olho seco foi definido como uma doença multifatorial da lágrima e da superfície ocular, resultando em desconforto, turvação visual e instabilidade do filme lacrimal, com dano potencial à superfície ocular. É associado ao aumento da osmolaridade da lágrima e à inflamação da superfície ocular.

O tratamento do olho seco deve levar em consideração a etiologia e a gravidade da doença. Medidas gerais para evitar a evaporação da lágrima, como evitar regiões secas e o uso de ar condicionado, são úteis, porém insuficientes. A terapêutica atual envolve, além de medidas comportamentais, a suplementação e retenção lacrimal, substitutos biológicos da lágrima, terapia anti-inflamatória e o uso de ácidos graxos essenciais [15].

As alterações oculares em pacientes infectados pelo HIV são comuns, afetando 70% a 80% destes indivíduos em algum ponto no curso da doença [16,17]. Um estudo [18] avaliou os principais sintomas da infecção pelo HIV em 4042 adultos. As queixas oftalmológicas, citadas por 32,4% dos pacientes, ficaram em oitavo lugar, o que reforça a importância do acompanhamento oftalmológico de indivíduos com este vírus.

A etiologia do olho seco associada ao vírus HIV ainda não é bem estabelecida. A redução lacrimal pode estar associada à infiltração linfocítica e eventual destruição dos ácinos e ductos da glândula lacrimal [19]. A prevalência do olho seco em pacientes infectados pelo HIV em estudos prévios variou de 7,79% a 38,8% e a sintomatologia do olho seco tem um impacto relevante na qualidade de vida destes pacientes [20-24].

O presente estudo avaliou 20 pacientes portadores do vírus HIV com diagnóstico de olho seco e comparou a eficácia do tratamento com suplementação lacrimal (carboximetilcelulose sódica 0,5%) ou com suplementação lacrimal associada à terapia anti-inflamatória (carboximetilcelulose sódica 0,5% e ciclosporina 0,05%).

MÉTODOS

Foi realizado um estudo prospectivo clínico randomizado de pacientes infectados pelo vírus da imunodeficiência humana (HIV) portadores de olho seco. Noventa e seis pacientes responderam ao questionário *Ocular Surface Disease Index* (OSDI®) e os que obtiveram pontuação maior que 25 e preencheram os critérios de inclusão e exclusão foram selecionados para exame oftalmológico completo e pesquisa de sinais de olho seco através dos seguintes testes: Tempo de ruptura do filme lacrimal (TRFL), teste de Schirmer I e coloração da superfície ocular por rosa bengala 1%. A realização dos testes foi rigorosamente semelhante para evitar fatores de confundimento.

O TRFL foi realizado instilando-se uma gota de fluoresceína sódica a 2% (Ophthalmos®, São Paulo, Brasil) no fundo de saco conjuntival inferior e solicitando ao paciente que piscasse várias vezes para, então, permanecer com o olho aberto. O observador, utilizando luz de cobalto na lâmpada de fenda, contou o tempo da abertura do olho até o aparecimento da mancha seca. Foi considerado compatível com olho seco quando menor ou igual a dez segundos [25].

A coloração da superfície ocular por rosa bengala foi realizada instilando-se uma gota de colírio rosa bengala a 1% (Ophthalmos®, São Paulo, Brasil) no fundo de saco conjuntival inferior e foi observado o tingimento conjuntival. Foi considerado compatível com olho seco quando ocorreu pontuação maior que três, segundo a escala de Van Bijsterveld [26].

O teste de Schirmer I foi realizado com tiras de papel filtro estéreis padronizadas (Ophthalmos®, São Paulo, Brasil), em que uma das extremidades foi dobrada na altura de cinco mm e inserida no fundo de saco conjuntival, na união do terço externo e médio da pálpebra inferior. Após 5 minutos, o papel filtro foi retirado e depois de 1 minuto de espera, foi medido o quanto ele se

umidificou. Foi considerado sugestivo de olho seco quando menor ou igual a 10 mm [27].

Foram incluídos 20 indivíduos infectados pelo vírus HIV, na faixa etária entre 18 e 45 anos que apresentaram sinais e sintomas de olho seco moderado. Foram excluídos os indivíduos que apresentassem outras condições associadas ao olho seco: coinfecção (vírus da hepatite B e C), mulheres na menopausa, portadores de doenças reumáticas e/ou em uso de medicações indutoras de olho seco (diuréticos, anti-histamínicos, betabloqueadores, antidepressivos, ansiolíticos), usuários de lente de contato, utilização de colírios beta bloqueadores e blefarite.

Os pacientes foram distribuídos aleatoriamente em dois grupos com 20 olhos cada e acompanhados durante seis meses. O grupo I foi tratado com colírio de carboximetilcelulose sódica 0,5% instilado quatro vezes ao dia, e o grupo II foi tratado com a associação de carboximetilcelulose sódica 0,5% (quatro vezes ao dia) e ciclosporina 0,05% tópica (duas vezes ao dia). Nas visitas, os pacientes responderam ao questionário OSDI® e foram submetidos ao TRFL, teste de Schirmer I e coloração por rosa bengala a 1%.

Foi comparada a eficácia do tratamento nos dois grupos através da curva de sobrevivência de Kaplan-Meyer e do teste Tarone-Ware, para as variáveis TRFL, Rosa Bengala e Schirmer I. Para comparar o comportamento dos grupos em relação ao OSDI®, os valores obtidos do questionário foram inicialmente testados para verificar a normalidade dos dados através do teste w/s e posteriormente verificados a diferença entre as variâncias usando o teste F, aplicando-se então o teste t para analisar a diferença de médias.

Resultados

A mediana de idade dos pacientes foi de 41 (32 a 45 anos), 11 (55%) eram mulheres, 9 (45%) eram homens, o menor tempo de diagnóstico de infecção pelo HIV foi 8 meses e o maior foi 20 anos. Os grupos I e II, formados por 10 pacientes cada, se mostraram homogêneos em relação ao sexo, idade e condições clínicas, satisfatoriamente controlados através da terapia HAART.

Pontuação no *Ocular Surface Disease Índex* (OSDI)®

Os valores mensurados pelo questionário OSDI® foram submetidos ao teste de normalidade w/s, no qual se obteve no tempo zero w/s 3.07 para o grupo I e 3.17 para o grupo II e no tempo 180 w/s 2.17 para o grupo I e 3.05 para o grupo II. Em geral, os dados parecem seguir uma distribuição normal.

O F-teste para comparar duas variâncias apresentou um p-valor de 0.44 para o tempo zero e p-valor de 0.78 para o tempo 180.

O t-teste para as duas amostras apresentou p-valor de 0.06 para o tempo zero e p-valor de 0.17 para o tempo 180.

Concluindo, não há significância estatística em relação aos grupos estudados com relação à variável OSDI®.

Tempo de ruptura do filme lacrimal

As curvas de sobrevivência de Kaplan-Meyer (Gráfico 1) indicam que os grupos não apresentaram diferença significativa entre si no decorrer dos 180 dias de tratamento, apesar do prolongamento do gráfico que representa um aumento no tempo de ruptura do filme lacrimal.

Coloração por rosa bengala

As curvas de sobrevivência de Kaplan-Meyer (Gráfico 2) indicam que os grupos não apresentaram diferença significativa entre si no decorrer dos 180 dias de tratamento, apesar da redução do gráfico que representa uma diminuição da pontuação da coloração da superfície ocular por rosa bengala.

Teste de Schirmer I

As curvas de sobrevivência de Kaplan-Meyer (Gráfico 3) indicam diferença estatisticamente significantes entre os dois grupos. Observa-se um prolongamento da curva contínua e uma manutenção da curva tracejada, o que representa uma maior produção lacrimal pelo grupo II.

Discussão

O olho seco é uma doença multifatorial da lágrima com dano potencial à superfície ocular, que resulta em desconforto, turvação visual e instabilidade do filme lacrimal [14], com significativo impacto na qualidade de vida [28,29].

Em nossa casuística, os grupos se mostraram homogêneos em relação ao sexo dos pacientes. Estudos prévios apresentaram resultados controversos em relação à prevalência da síndrome do olho seco em homens e mulheres portadores de HIV. Foram descritas prevalências iguais a 21% [24] e 38.8% [21] em homens e 16.9% [23] em mulheres. Rodrigues et al. [30] relataram prevalência igual a 25,8%, sem diferença em relação ao sexo. Na população geral, o sexo feminino é considerado um fator de risco consistente para desenvolvimento de olho seco [3,28]. Estas diferentes prevalências podem estar relacionadas aos critérios diagnósticos para olho seco

Gráfico 1

Curva de sobrevivência de Kaplan-Meyer e teste de Tarone-Ware para o tempo de ruptura do filme lacrimal

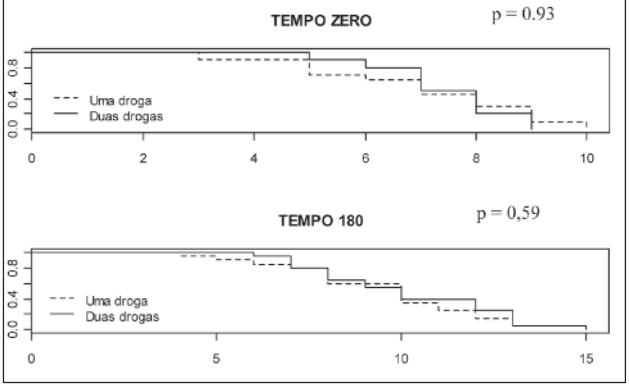


Gráfico 2

Curva de sobrevivência de Kaplan-Meyer e teste de Tarone-Ware para a coloração por rosa bengala

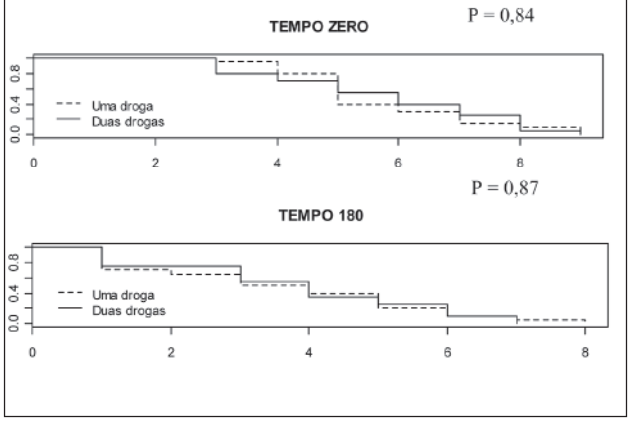
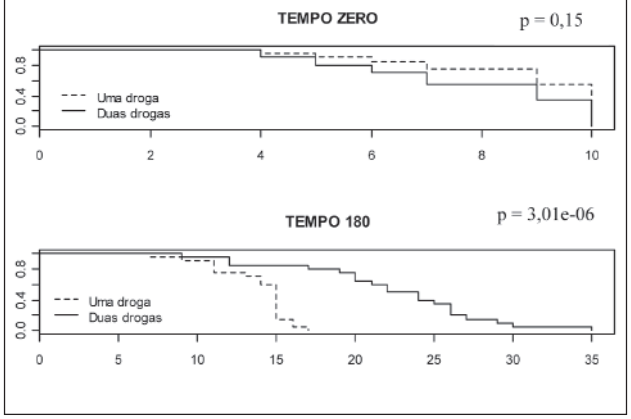


Gráfico 3

Curva de sobrevivência de Kaplan-Meyer e teste de Tarone-Ware para o teste de Schirmer I



utilizados, ao número de pacientes avaliados e à idade dos pacientes.

No presente estudo, foram excluídos pacientes com idade superior a 45 anos. A idade é um fator de risco bem definido para olho seco [12,28]. A secreção lacrimal começa a declinar por volta dos 30 anos; entretanto, a produção excede a necessidade basal e não há sintomatologia. O nível crítico é atingido por volta de 45 anos, quando, em algumas circunstâncias, há sintomas relacionados ao olho seco. Aos 60 anos, a maioria das pessoas queixa-se de olho seco em algumas situações, tais como: exposição ao ar condicionado, uso de lentes de contato e à noite, quando há menor produção de lágrima devido à variação cicardiana [12].

Em nossa casuística, os grupos se mostraram homogêneos em relação ao tempo de diagnóstico da infecção pelo HIV. Rodrigues e cols. [30] observaram um aumento significativo da freqüência da síndrome de olho seco nos pacientes com doença acima de quatro anos, principalmente entre os que estavam em uso de terapia antirretroviral oral. No presente estudo, os pacientes apresentaram a doença controlada em uso de terapia HAART e os grupos eram homogêneos em relação às condições clínicas. Rodrigues e cols. [30] não encontraram relação entre o decréscimo da produção lacrimal e a gravidade da infecção pelo vírus HIV. Geier et al. [22] também não encontraram associação entre a contagem de CD4 ou a gravidade da infecção pelo HIV com a síndrome do olho seco.

Os principais estudos epidemiológicos [2-4,7-10] sobre a síndrome do olho seco determinaram a prevalência de acordo com os sintomas referidos, tais como: ressecamento, sensação de areia, queimação, hiperemia, olhos colados pela manhã, prurido e fotofobia. Tais sintomas são identificados através da aplicação de questionários como o OSDI®.

Os questionários são empregados para rastreamento de indivíduos com olho seco, para verificar o impacto do tratamento nos sintomas ou para graduar a gravidade da doença [28]. O OSDI® é um questionário que avalia a gravidade do olho seco através de 12 perguntas que abordam sinais e sintomas, dificuldades visuais e fatores desencadeantes na semana anterior à avaliação. Este teste foi escolhido por ser rápido, confiável e de fácil entendimento, resultando em uma

pontuação mesmo que o paciente não tenha respondido a todas as perguntas [31-33].

O presente estudo avaliou subjetivamente os sintomas e objetivamente os sinais relacionados ao olho seco. Foram utilizados, além do questionário OSDI®, três testes diagnósticos para olho seco (TRFL, teste de Schirmer I e coloração por rosa bengala a 1%). Não existe um teste padrão-ouro para o diagnóstico de olho seco [34]. Entretanto, estes testes são bem estabelecidos na propedêutica desta síndrome. Segundo o painel Delphi [13], os testes mais frequentemente utilizados pelos especialistas são a coloração por fluoresceína, o TRFL, o teste de Schirmer e a coloração por rosa bengala. Estudos prévios avaliaram a sensibilidade e especificidade destes testes. Vitali et al. [35] encontraram sensibilidade igual a 83% e especificidade igual a 68% para o valor de referência de 10mm em cinco minutos para o teste de Schirmer I. Os mesmos autores encontraram sensibilidade igual a 72% e especificidade igual a 62% para o valor de referência menor que 10 segundos para o TRFL. Goren et al. [36] encontraram sensibilidade igual a 25% e especificidade igual a 90% para qualquer pontuação na coloração da superfície ocular com rosa bengala.

O objetivo do tratamento do olho seco é melhorar os sintomas e os sinais da superfície ocular e a escolha do tratamento deve considerar a etiologia e a gravidade da doença.

As lágrimas artificiais lubrificam e hidratam a superfície ocular através da substância ativa da sua composição. Entretanto, os conservantes utilizados para manter a lágrima artificial estéril podem danificar o epitélio dependendo da frequência de uso do colírio. [37,38] No presente estudo, foi utilizada a substância ativa carboximetilcelulose sódica na concentração de 0,5% com o conservante clorito de sódio, que quando exposto à luz, se decompõe em íons cloreto e água, evitando o dano epitelial devido ao uso prolongado do colírio [37,38].

Pacientes com olho seco possuem inflamação na superfície ocular e nas glândulas lacrimais mediadas por citocinas [14]. A ciclosporina é uma droga imunomoduladora que inibe a ativação das células T, a produção destas citocinas inflamatórias e a cascata da apoptose [39]. A eficácia da ciclosporina no tratamento do olho seco já foi comprovada em estudos prévios [40-42], porém não há relatos do seu uso no olho seco associado à infecção pelo HIV.

Com relação ao questionário OSDI®, aplicado em todas as visitas, observa-se que nos dois grupos há redução da pontuação, o que indica melhora dos sintomas, todavia não ocorreu diferença estatisticamente significante entre os dois grupos ao final do tratamento.

Em nossa casuística, embora o TRFL tenha aumentado nos dois grupos, não houve uma diferença estatisticamente significante do TRFL entre os grupos após 180 dias de tratamento.

Com relação à coloração da superfície ocular por rosa bengala, houve redução da pontuação de acordo com a escala de van Bijsterveld nos grupos I e II, porém a diferença entre os dois grupos não foi estatisticamente significante ao término do tratamento.

Com relação ao teste de Schirmer I, ocorreu um aumento dos valores após o tratamento nos dois grupos, entretanto o tratamento do grupo II mostrou-se superior ao tratamento do grupo I.

O efeito benéfico da ciclosporina na produção lacrimal é bem estabelecido. Quando a etiologia do olho seco relaciona-se à diminuição da produção aquosa e à inflamação da superfície ocular, sabe-se que a ciclosporina é bem indicada. Espera-se que o aumento da produção lacrimal resulte em redução da sintomatologia e dos sinais na superfície ocular. Entretanto, o grupo II não apresentou diferenças nos sintomas e nos outros sinais, apesar da melhora significativa do teste de Schirmer I em relação ao grupo I. Tal fato reforça o conceito de que o olho seco é uma doença multifatorial e ainda não completamente elucidada em pacientes HIV-positivos.

A literatura carece de estudos que avaliaram o tratamento de pacientes infectados pelo HIV com olho seco. O presente estudo é o primeiro trabalho que avaliou a eficácia do tratamento do olho seco em pacientes portadores do vírus HIV com suplementação lacrimal e terapia anti-inflamatória e sugere que o uso de ciclosporina tópica melhore a produção lacrimal nestes pacientes. Entretanto, é possível que o tratamento destes pacientes devesse envolver medidas adicionais. O papel do HIV na patogênese do olho seco deve ser mais bem investigado para direcionar melhor a terapêutica.

Algumas limitações deste estudo merecem considerações. Seria interessante realizar o estudo com um número maior de pacientes, para melhor avaliação do impacto do tratamento na sintomatologia e no exame oftalmológico dos portadores do vírus HIV e olho seco. Como o olho seco é uma doença multifatorial, qualquer outra condição associada ao olho seco foi considerada critério de exclusão, para evitar fatores de confundimento. Além disso, para aumentar a adesão ao tratamento, o presente estudo não incluiu pacientes com sintomas de olho seco leve. Tais critérios metodológicos reduziram a amostra utilizada. A exata fisiopatologia do olho seco em pacientes infectados pelo HIV ainda não é totalmente

compreendida e a definição da síndrome do olho seco ainda é controversa, o que dificulta o diagnóstico desta doença complexa.

Conclusão

O uso de carboximetilcelulose sódica 0,5% foi tão eficaz quanto o uso da carboximetilcelulose sódica 0,5% associada à ciclosporina 0,05% na redução dos sintomas medidos pelo OSDI®, assim como na melhora do TRFL e da coloração da superfície ocular por rosa bengala a 1% em pacientes infectados pelo HIV com olho seco.

Entretanto, comparativamente, o tratamento do olho seco com suplementação lacrimal (carboximetilcelulose sódica 0,5%) associada à terapia anti-inflamatória (ciclosporina 0,05% tópica) mostrou-se superior ao uso isolado de carboximetilcelulose sódica 0,5% em relação à produção lacrimal, medida pelo teste de Schirmer I.

Abstract

Purpose: *This study evaluates the use of topical cyclosporine 0.05% and sodium carboxymethylcellulose 0.5% for the treatment of dry eye disease in patients infected with the human immunodeficiency virus.* ***Methods:*** *Twenty HIV-positive-patients were selected from the Department of Infectious Diseases of the Federal University of Rio de Janeiro Hospital. Dry eye diagnosis was based on a dry eye questionnaire (Ocular Surface Disease Index - OSDI®), Schirmer I test, break up time and 1% rose bengal staining of the ocular surface. The patients were divided into two groups with ten patients (20 eyes). Group I received sodium carboxymethylcellulose 0.5% drops and group II received sodium carboxymethylcellulose 0.5% drops and topical cyclosporine 0.05% for six months.* ***Results:*** *Comparing the results at the beginning and in the end of the treatment, the use of cyclosporine was superior ($p < 0.05$) for the Schirmer I test.* ***Conclusion:*** *This study suggests that topical cyclosporine improves lacrimal production and has an important role as an adjuvant therapy for dry eye disease in patients infected with the human immunodeficiency virus.*

ClinicalTrials.gov Identifier: NCT00797030

Keywords: *Dry eye syndromes/drug therapy; keratoconjuntivitis sicca/drug therapy, Cyclosporine/therapeutic use; Ophthalmic solutions/administration & dosage; Administration, topical; HIV seropositivity*

Referências

1. Johnson ME, Murphy PJ. Changes in the tear film and ocular surface from dry eye syndrome. Progress in retinal and eye research 2004;23(4):449-74.
2. McCarty CA, Bansal AK, Liv ingston PM, Stanislavsky YL, Taylor HR. The epidemiology of dry eye in Melbourne, Australia. Ophthalmology. 1998;105(6):1114-9.
3. Schaumberg DA, Sullivan DA, Buring JE, Dana MR. Prevalence of dry eye syndrome among US women. Am J Ophthalmol. 2003:136(2);318-26.
4. Moss SE, Klein R, Klein BE. Prevalence of and risk factors for dry eye syndrome. Arch Ophthalmol. 2000;118(9):1264-8.
5. Schein OD, Hochberg MC, Munoz B, Tielsch JM, Bandeen-Roche K, Provost T, et al. Dry eye and dry mouth in the elderly: a population-based assessment. Arch Intern Med. 1999;159(12):1359-63.
6. Schein OD, Munoz B, Tielsch JM, Bandeen-Roche K, West S. Prevalence of dry eye among the elderly. Am J Ophthalmol. 1997;124(6):723-8.
7. Munoz B, West SK, Rubin GS, Schein OD, Quigley HA, Bressler SB , et al. Causes of blindness and visual impairment in a population of older Americans: The Salisbury Eye Evaluation Study. Arch Ophthalmol. 2000;118(6):819-25.
8. Chia EM, Mitchell P, Rochtchina E, Lee AJ, Maroun R, Wang JJ. Prevalence and associations of dry eye syndrome in an older population: the Blue Mountains Eye Study. Clin Experiment Ophthalmol. 2003;31(3):229-32.
9. Lee AJ, Lee J, Saw SM, Gazzard G, Koh D, Widjaja D, et al. Prevalence and risk factors associated with dry eye symptoms: a population based study in Indonesia. Br J Ophthalmol. 2002;86(12):1347-51.
10. Lin PY , Tsai SY , Cheng CY , Liu JH, Chou P, Hsu WM. Prevalence of dry eye among an elderly Chinese population in Taiwan: the Shihpai Eye Study. Ophthalmology. 2003;110(6):1096-101.
11. Lemp MA. Report of the National Eye Institute/Industry workshop on Clinical Trials in dry eyes. CLAO J. 1995;21(4):221-32.
12. Murube J, Benitez Del Castillo JM, Chenzhuo L, Berta A, Rolando M. Triple clasificación de Madrid para el ojo seco. Arch Soc Esp Oftalmol. 2003;78(11):587-93;595-601.
13. Behrens A, Doyle JJ, Stern L, Chuck RS, Mc Donnell PJ, Azar DT, et al. Dysfunctional tear syndrome: a Delphi approach to treatment recommendations. Cornea. 2006;25(8): 900- 7.
14 The definition and classification of dry eye disease: report of the definition and Classification Subcommittee of the International Dry Eye Workshop (2007). Ocul Surf 2007;5(2):75-92.
15. Management and therapy of dry eye disease: report of the Management and Therapy Subcommittee of the International Dry Eye Workshop (2007). Ocul Surf. 2007;5(2):163-78.
16. Cunningham ET Jr. Uveitis in HIV positive patients. Br J Ophthalmol 2000;84(3):233-7.
17. Cunningham ET Jr, Margolis TP. Ocular manifestations of HIV infection. N Engl J Med. 1998;339(4):236-44.
18. Mathews WC, McCutchan JA, Asch S, Turner BJ, Gifford AL, Kuromiya K, et al. National estimates of HIV-related symptom prevalence from the HIV cost and service utilization study. Med Care. 2000;38(7):750-62.
19. Pflugfelder SC, Wilhelmus KR, Osato MS, Matoba AY, Font RL. The autoimmune nature of aqueous tear deficiency. Ophthalmology 1986;93(12):1513-7.

20. Kordossis T, Paikos S, Aroni K, Kitsanta P, Dimitrakopoulos A, Kavouklis E, et al. Prevalence of Sjögren-like syndrome in a cohort of HIV-1 positive patients: descriptive pathology and immunopathology. Br J Rheumatol. 1998:37(6):691-5.
21. DeCarlo DK, Penner SL, Schamerloh RJ, Fullard RJ. Dry eye among males infected with the human immunodeficiency virus. J Am Optom Assoc. 1995;66(9):533-8.
22. Geier SA, Libera S, Klauss V, Goebel FD. Sicca syndrome in patients infected with the human immunodeficiency virus. Ophthalmology. 1995;102(9):1319-24.
23. Lucca JA, Kung JS, Farris RL: Keratoconjunctivitis sicca in female patients infected with human immunodeficiency virus. CLAO J. 1994;20(1):49-51.
24. Lucca JA, Farris RL, Bielory L, Caputo AR. Keratoconjunctivitis sicca in male patients infected with human immunodeficiency virus type 1. Ophthalmology.1990;9798):1008—10.
25. Yamane R, Yamane Y. Testes de secreção e drenagem lacrimal e estesiometria. In: Yamane R. Semiologia ocular. 2a ed. Rio de Janeiro: Cultura Médica; 2003 p.83-88
26. Van Bijsterveld OP. Diagnostic tests in the sicca syndrome. Arch Ophthalmol. 1969;82(1):10-4.
27. Mattar DB, Kara-José N. Filme lacrimal. In: Belfort Jr, R, Kara-José, N. Córnea clínica – cirúrgica. São Paulo: Roca; 1996: 33-40.
28. The epidemiology of dry eye disease: report of the Epidemiology Subcommittee of the International Dry Eye WorkShop (2007). Ocul Surf. 2007;5(2):93-107.
29. Miljanovic B, Dana R, Sullivan DA, Schaumberg DA. Impact of dry eye syndrome on vision-related quality of life. Am J Ophthalmol. 2007;143(3):409-15.
30. Rodrigues ML, Rodrigues MLV, Holanda de Freitas JA. Estudo da síndrome de ceratoconjuntivite seca de pacientes soropositivos para o vírus da imunodeficiência adquirida humana tipo 1 e com síndrome da imunodeficiência adquirida, em uso ou não de terapia anti-retroviral combinada (HAART). Arq Bras Oftalmol. 2004;67(2):283-7.
31. Schiffman RM, Christianson MD, Jacobsen G, Hirsch JD, Reis BL. Reliability and validity of the Ocular Surface Disease Index. Arch Ophthalmol. 2000;118(5):615-21.
32. Ozcura F, Aydin S, Helvaci MR. Ocular Surface Disease Index for the diagnosis of dry eye syndrome. Ocul Immunol Inflam 2007;15(5):389-93.
33. Vitale S, Goodman LA, Reed GF, Smith JA. Comparison of the NEI-VFQ and OSDI questionnaires in patients with Sjogren's syndrome-related dry eye. Health Qual Life Outcomes. 2004;2:44.
34. Methodologies to diagnose and monitor dry eye disease: report of the Diagnostic Methodology Subcommittee of the International Dry Eye WorkShop (2007).Ocul Surf. 2007;5(2):108-52.
35. Vitali C, Moutsopoulos HM, Bombardieri S. The European Community Study Group on diagnostic criteria for Sjögren´s syndrome. Sensitivity and specificity of tests for ocular and oral involvement in Sjögren´s syndrome. Ann Rheum Dis. 1994;53(10):637-47.
36. Goren MB, Goren SB. Diagnostic tests in patients with symptoms of keratoconjunctivitis sicca in clinical practice. Am J Ophthalmol. 1988;106(5):570-4.
37. Management and therapy of dry eye disease: report of the Management and Therapy Subcommittee of the International Dry Eye WorkShop (2007). Ocul Surf. 2007;5(2):163-78.
38. Calonge M. The treatment of dry eye. Surv Ophthalmol. 2001;45 Suppl 2:227-39.
39. Nussenblatt RB, Palestine AG. Cyclosporin: immunology, pharmacology and therapeutic uses. Surv Ophthalmol. 1986;31(3):159-69.
40. Fullard RJ, Kaswan RM, Bounous DI, Hirsh SG. Tear protein profiles VS. Clinical characteristics of untreated and cyclosporine-treated canine KCS. J Am Optom Assoc. 1995;66(7):397-404.
41. Tsubota K, Saito I, Ishimaru N, Hayashi Y: Use of topical cyclosporine A in a primary Sjogrens syndrome mouse model. Invest Ophthalmol Vis Sci. 1998;39(9):1551–9.
42. Perry HD, Solomon R, Donnenfeld ED, Perry AR, Wittpenn JR, Greenman HE, et al. Evaluation of topical cyclosporine for the treatment of Dry Eye Disease. Arch Ophthalmol. 2008;126(8):1046-50.

ENDEREÇO PARA CORRESPONDÊNCIA:
Rodrigo de Pinho Paes Barreto
Rua Vilhena de Moraes, nº 100 - bloco 02 - apt 903
CEP 22793-140 - Barra da Tijuca - RJ
E-mail: barreto@live.com

Apêndice D

Algoritmo EM e Critério de Informação de Akaike

O Algoritmo EM (Seção 2.4.2) e o critério de Informação de Akaike (Seção 2.4.3) foram utilizados para estudar a antigenicidade plaquetária do subgrupo sanguíneo A em indivíduos euro e afro-brasileiros:

SANT'ANNA GOMES, B. M., ESTALOTE, A. C., PALATNIK, M., PIMENTA, G., PEREIRA, B. DE B., NASCIMENTO, E. M. Prevalence, serologic and genetic studies of high expressers of the blood group A antigen on platelets, *Transfusion Medicine*, v. 20, pp. 303-314, 2010.

Transfusion Medicine
Official Journal of the British Blood Transfusion Society

Transfusion Medicine, 2010, **20**, 303–314 doi: 10.1111/j.1365-3148.2010.01017.x

ORIGINAL ARTICLE

Prevalence, serologic and genetic studies of high expressers of the blood group A antigen on platelets*

B. M. Sant'Anna Gomes,[1] A. C. Estalote,[1] M. Palatnik,[1,2] G. Pimenta,[3] B. de B. Pereira[4,5] & E. M. do Nascimento[5] [1]*Blood Transfusion Service Research Laboratory, Hospital Universitário Clementino Fraga Filho (HUCFF),* [2]*Postgraduate Program on Clinical Medicine (Hematology), Faculty of Medicine,* [3]*Division of Hematology, HUCFF,* [4]*Scientific Investigation Committee, HUCFF, Faculty of Medicine, and* [5]*Postgraduate Programs in Engineering (COPPE), Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro (RJ), Brazil*

Received 16 April 2009; accepted for publication 7 May 2010

SUMMARY

Objective/Aim: The aim of this study is to describe the distribution of the platelet blood group A antigenicity in Euro-Brazilians (EUBs) and Afro-Brazilians (AFBs).

Background: A small but significant proportion of individuals express high levels of A or B antigen on their platelets corresponding to the erythrocyte ABO group. The mechanism of increased antigen expression has not been elucidated.

Material/Methods: A cohort of 241 blood group A donors was analysed by flow cytometry. Although mean fluorescence intensity (MFI) is a typical continuous variable, platelets were screened and divided into two categories: low expressers (LEs) and high expressers (HEs). A three-generation family was investigated looking for an inheritance mechanism.

Results: The prevalence of the HE platelet phenotype among group A_1 donors was 2%. The mean of MFI on platelets of A_1 subgroup of EUBs differs from that of AFBs ($P = 0{\cdot}0115$), whereas the frequency of the HE phenotype was similar between them ($P = 0{\cdot}5251$). A significant difference was found between sexes ($P = 0{\cdot}0039$). Whereas the serum glycosyltransferase from HE family members converted significantly more H antigen on group O erythrocytes into A antigens compared with that in LE serum, their *ABO*, *FUT1* and *FUT2* genes were consensus. The theoretically favourable, transcriptionally four-repeat *ABO* enhancer was not observed.

Conclusion: The occurrence of HE in several members suggests familial aggregation. Indeed, in repeated measures, stability of the MFI values is suggesting an inherited condition. Factors outside the *ABO* locus might be responsible for the HE phenotype. Whether the real mechanism of inheritance is either of a polygenic or of a discrete Mendelian nature remains to be elucidated.

Key words: Platelets, ABO blood group, *ABO* gene, Genetic Polymorphism, High Expresser Platelet Phenotype.

Correspondence: Prof. Marcos Palatnik, Rua Lauro Muller 56/203, CEP 22290-160 Botafogo, Rio de Janeiro, Brazil.
Tel.: +55 (21) 3820-9224; fax: +55 (21) 2562-2460;
e-mail: palatos@infolink.com.br

*Part of this work was presented at the 28th Congress of the Brazilian Society of Hematology and Blood Transfusion held in Rio de Janeiro, Brazil, November 2005, and at the 58th Annual Meeting of the American Association of Blood Banks held in Seattle, WA, USA, October 2005, published in abstract forms [Sant'Anna Gomes B.M. *et al*. (2005) *Brazil J Hematol Hemother*, **27** (Suppl. 2), 296 and 304–5; Yazer M.H. *et al*. (2005) *Transfusion*, **45** (Suppl. 3), 130A].

Human platelets express alloantigens that are platelet specific (e.g. Human Platelet Antigens (HPA)) and others that are shared with other blood cells and tissues (e.g. ABH, P, Le, I and Human Leucocyte Antigens (HLA) class I determinants) (Kunicki & George, 1989). Although it has been known for decades that human platelets express A and B antigens corresponding to the ABO blood group of the individual's erythrocytes, many blood banks worldwide transfuse platelets without regarding donor–recipient ABO compatibility. Although this practice might relieve inventory pressures, several reports have described platelet refractoriness mediated

by anti-A and anti-B antibodies (Skogen *et al*., 1988; Ogasawara *et al*., 1993; Curtis *et al*., 2000).

In Japan, approximately 7% of the population expresses significantly elevated levels of either A or B antigen on the platelet surface (Ogasawara *et al*., 1993). It has been suggested that the expression of ABH antigens on platelets is genetically determined and that individuals can be classified as either low expressers (LEs) or high expressers (HEs) depending on the amount of A or B antigen detectable on their platelets. Curtis *et al*. (2000) showed that platelet A antigen levels in 7% of Caucasian blood group A_1 donors were higher than in the general population mean + 2SD using flow cytometry. The corresponding percentage for B antigen in B donors was 4%. Furthermore, these investigators proposed that HE individuals could be divided into two categories of HEs of group A or B antigen, based on the flow cytometrically generated histogram pattern. The amount of these antigens on the platelets in the HE type I (HE-I) group overlapped with the mean of that in the general population but was, on average, higher. In the HE type II (HE-II) group, there was a clear separation between the amount of platelet A (or B) antigen on these donors' platelets and the mean of the general population as shown by the histogram.

A recent study indicated that individuals with the A_2 red blood cell (RBC) phenotype may lack A antigen as well as its precursor substance H antigen on their platelets (Cooling *et al*., 2005), although this finding has been questioned by others (Curtis & Aster, 2006). We have evaluated the amount of A antigen on the platelet surface of 241 Brazilian blood group A_1, A_2 and A_{int} donors. This is the largest cohort of A donors studied so far and also the first report about ethnic groups. We also present a study on a group A_1 HE Afro-Brazilian (AFB) family with a complete analysis of the underlying *ABO* gene sequence including its 5′ enhancer region, suggested by some investigators (Gershan *et al*., 2006) to play a role in *ABO* transcriptional platelet regulation. The *FUT1* and *FUT2* genes, which are involved in synthesising the ABO precursor substance (H antigen) on RBCs and in secretions, respectively, have also been investigated. The occurrence of HE phenotype in several members suggests familial aggregation of the character. The potential mechanism of inheritance of the platelet HE phenotype is also discussed.

MATERIALS AND METHODS

Specimens

On giving blood, 241 Brazilian group A blood donors agreed to participate in the study. The study was approved by the Ethical Scientific Committee (Comitê de Ética em Pesquisa, registration number: 001/003, Hospital Universitário Clementino Fraga Filho/ Faculdade de Medicina, Universidade Federal do Rio de Janeiro, Brazil), and all subjects received oral and written information concerning the study and gave their informed consent. The ethnicity of the donors, either Euro-Brazilian (EUB) or AFB, was assigned at the time of blood donation, as previously reported (Palatnik *et al*., 2002). Furthermore, a possible bone marrow donor, who had shown the highest level of platelet A antigen, and his 13 living relatives were also analysed. Informed consent was obtained from each individual or from the parents on behalf of their children. All of the family members were AFBs born in Rio de Janeiro, Brazil.

Following Michelson *et al*. (2000), for platelet studies, a 4-mL venous blood sample was collected from each individual into 3·2% sodium citrate-coated polyethylene terephthalate tubes (Vacuette®, Greiner Bio-One, Americana, SP, Brazil). Saliva and additional blood samples (8 mL in ethylenediaminetetraacetic acid and 10 mL in tubes without additives) from family members and some blood donors for platelet ABO antigen reproducibility tests, glycosyltransferase (GTA) assays and genetic testing were also obtained as described in subsequent sections.

Erythrocyte phenotyping

The donors' blood groups were initially determined by slide test with monoclonal anti-A and anti-B reagents (DiaMed, Lagoa Santa, MG, Brazil) and repeated using the microplate system (DiaMed-MP Test®, DiaMed, Cressier sur Morat, Switzerland). To differentiate between the A subgroups, anti-A_1 (*Dolichos biflorus*, Biotest, Itapecerica da Serra, SP, Brazil, and mouse hybridoma, Biotest, Dreieich, Germany) and anti-H (our own crude extracts of *Ulex europaeus* seeds and a commercial lectin from Biotest, Brazil) testing were carried out according to the manufacturers' instructions. The definitions of the A subgroups followed published recommendations (Palatnik, 1984; see also Daniels, 2002).

Platelet preparation

To avoid *ex vivo* platelet activation, samples were processed within 30 min of blood drawing for all assays (Mody *et al*., 1999; Michelson, 2006). Each blood sample was centrifuged at 22 °C for 5 min at $500 \times g$ in a Jouan Model BR 4i centrifuge (Societe Jouan, St Herblain, France). To minimise the formation of platelet aggregates, one part of the plasma supernatant (platelet-rich plasma, PRP) was diluted in two parts of 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)-buffered saline [0·14 M NaCl, 5 mM KCl,

1 mM $MgSO_4$ and 10 mM HEPES (sodium salt), pH 7·4]. Samples were stabilised by fixation in paraformaldehyde at a final concentration of 1% (Metcalfe *et al*., 1997; Mody *et al*., 1999; Michelson *et al*., 2000; Michelson, 2006).

Quantitative flow cytometric assay

Fixed platelets (50 μL of PRP aliquot, containing approximately 8×10^7 μL^{-1} platelets) were incubated undisturbed for 30 min in the dark at room temperature with saturating dilutions (predetermined by titrations) of fluorescein isothiocyanate- (FITC)- and R-phycoerythrin- (R-PE, PE)-labelled monoclonal antibodies. The antibodies used for dual-colour staining were (i) anti-CD41 (mouse anti-human CD41, platelet glycoprotein IIb/IIIa R-PE, Dako Cytomation, Glostrup, Denmark) and anti-A antibody [FITC-conjugated mouse $IgG_3\kappa$ anti-A monoclonal antibody, clone NaM87-1F6 (Becton-Dickinson Co., BD Pharmingen, San Diego, CA, USA)] and (ii) anti-CD41 and anti-H lectin (lectin, FITC labelled, from *U. europaeus* UEA I, Sigma Chemical Co., St Louis, MO, USA). The quantity of antibody used was sufficient to saturate available A antigen sites: some assays when repeated by incubating the conjugate supernatant with a new platelet aliquot yielded histograms similar to the initial ones.

In order to control the *in vitro* activation, fixed platelets were incubated with anti-CD62p (PE-conjugated mouse anti-human CD62p monoclonal antibody, BD Pharmingen). Any possible neutrophil contamination was discarded by incubation with anti-CD45 (R-PE monoclonal mouse anti-human CD45, leucocyte common antigen, Dako). To detect nonspecific antibody binding, respective isotype controls (negative controls: R-PE-conjugated IgG_1 mouse monoclonal antibody, Dako; PE-conjugated IgG_1 mouse monoclonal antibody, BD Pharmingen; and FITC-conjugated mouse $IgG_3\kappa$, BD Pharmingen) were also used under the same incubation conditions described earlier.

After staining, platelets were washed ($500 \times g$, 22 °C, 5 min) twice in 0·1 M phosphate-buffered saline (PBS), 0·02% NaN_3 (w/v), 0·1% bovine serum albumin (BSA) (w/v), pH 7·2–7·4, and resuspended in 500 μL of wash solution (Metcalfe *et al*., 1997; Mody *et al*., 1999; Michelson *et al*., 2000; Michelson, 2006). Platelet-bound fluorescence and control samples (50 μL aliquot of fixed platelets from the same individual) were analysed within 2 h by flow cytometry (FACSCalibur, BD Pharmingen). Forward scatter vs side scatter and forward scatter vs fluorescence gates (FL) of platelets were set. A total of 30 000 events per sample were gated and analysed with the computer software recommended by the manufacturer (CellQuest™ Software, BD Pharmingen).

Results were recorded as the mean fluorescence intensity (MFI) of the platelets positive for both CD41 and either anti-A or anti-H reagents. Maximal platelet activation was considered in the order of ≤1·5% (CD62p expression).

One-colour fluorescence flow cytometry was performed on the RBCs of some individuals. One part of washed RBCs was diluted in two parts of HEPES/BSA-buffered saline and stabilised by fixation in paraformaldehyde at a final concentration of 1% (Metcalfe *et al*., 1997; Mody *et al*., 1999; Michelson *et al*., 2000; Michelson, 2006). An aliquot (5 μL) of fixed RBCs was stained with either anti-A antibody or anti-H lectin. After staining, RBCs were washed twice and resuspended in wash solution before flow cytometry analysis. The results were reported as the MFI in relation to each fluorochrome.

Serum enzyme activity assays

The 3-α-*N*-acetylgalactosaminyltransferase (blood group A GTA) activity was measured in three A_1 blood group HE family members following the method of Yabe *et al*. (1989), using serum from five A_1 blood group LE individuals for comparison. In order to assay the GTA activity, 200 μL of 1·6 mM UDP-*N*-acetylgalactosamine (Sigma), 400 μL of 0·1 M cacodylic acid, pH 6·0 (SPI Supplies, West Chester, PA, USA), containing 0·1 M $MnCl_2$ (Sigma), and 100 μL of 50% suspension of group O RBCs treated with 0·1% papain solution (Sigma, lot: 127F8075) were added to 1 mL of each serum. The mixture was agitated at 37 °C for 1 h. After incubation, the cells were washed three times in a solution of 0·1 M PBS and 0·5% BSA. A fraction of the converted RBCs was titrated against human anti-A serum. The remainder, after paraformaldehyde fixation, as described earlier, was incubated with FITC-anti-A antibody and tested by flow cytometry.

Saliva secretion assays

Saliva samples from family members collected in sterile tubes were immediately inactivated by immersion in boiling water for 15 min. After centrifugation ($2400 \times g$, 10 min), quantitative saliva inhibition tests were performed, as published elsewhere (Mallory, 1993).

Molecular analysis of the ABO and FUT genes

Some members of the AFB family were subjected to molecular analysis. An initial *ABO* genotyping screen was performed using two independent methods: polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) and PCR-allele-specific primer

(ASP) (Olsson & Chester, 1995, 1996; Olsson *et al*., 1998, 2001). Direct DNA sequencing of exons 1–7, as previously described (Olsson *et al*., 2001), was performed in two samples from family members. In addition, enhancer repeat PCR was used to determine the allele-dependent number of CCAAT-binding factors (CBF/NF-Y) motifs found in this minisatellite region (approximately 4 kbp upstream of the ABO start codon in exon 1) (Irshaid *et al*., 1999).

The coding regions of the *FUT1* and *FUT2* genes were amplified as one fragment per locus (1171 and 1188 bp, respectively) using the following primers: FUT1·1F, ctcagcctcagagcatttgc; FUT1·3R, gctacttcagaaagtctccctg; FUT2-43F, ccatctcccagctaacgtgtcc; FUT2+43R, gggaggcagagaaggagaaaagg (Yip *et al*., 2002; Storry *et al*., 2004). For sequence analysis, the amplified DNA bands were eluted from the gel using the QIAquick gel purification kit (Qiagen Nordic, Crawley, UK) and sequenced directly using an ABI 3130 sequencer (Applied Biosystems, Inc., Foster City, CA, USA) and BigDye® reagents (Applied Biosystems).

Statistical analysis

Log-transformed (natural logarithms, ln) continuous variables were used, where appropriate. Furthermore, when necessary, the log values were converted to MFI values. The ordinates of the graphics express the expected values for $M = e^{(\mu+VAR/2)}$, and $SEM = [(\sqrt{(e^{VAR} - 1)(e^{2\mu+VAR})}]/\sqrt{(n-1)}$, which are based on the ln (μ and VAR) of the original values read by flow cytometry, where M is the mean MFI value, VAR is the variance and SEM is the standard error of the mean. Groups were compared by Kruskal–Wallis and Mann–Whitney *U* tests. Association between variables was assessed by Spearman's correlation coefficient. Linear regression analysis was used to assess the strength and independence of associations between variables. Categorical variables were compared by χ^2 and Fisher's exact tests. Normality of the distributions was checked by Kolmogorov–Smirnov test. Stability of the MFI values was checked by one-way ANOVA for three correlated samples. This analysis involves each subject being measured in each of the *k* conditions (repeated measures or within-subjects design). A single normal distribution vs two normal mixture or three normal mixture distributions were checked by expectation-maximization (EM) algorithm and by Akaike's information criteria (AIC) (Pawitan, 2001; Burnham & Anderson, 2002). Statistical analysis was performed using SPSS (version 10·0 for Windows), ANALYSE-IT for Microsoft Excel (version 2·07) and VASSARSTATS software. Differences were considered significant with P values $<0{\cdot}05$.

RESULTS

The mean and SD of the population of individual values of MFI were used to define the two empirical categories of A antigen expression on platelets among A_1 blood group donors: LE, values lower or equal to the mean + 1·96 SD; HE, values higher than the mean + 1·96 SD.

Samples from 241 blood group A donors ($A_1 = 183$, $A_{int} = 17$ and $A_2 = 41$) and 50 blood group O donors were obtained. The ln mean + 1·96 SD = 4·82 was used to classify the A_1 phenotypes: LE = 3·86 ± 0·46 and HE = 5·02 ± 0·17. The prevalence of platelet HE phenotype was estimated in 2% of the A_1 blood donors.

Ethnic, age and sex distribution

Differences in the amount of A antigen expressed on platelets of the A_1 donors between ethnic and sex subsamples of the population were found to be significant (Fig. 1). There was no correlation between MFI and age for the A_1 individuals ($P = 0{\cdot}5947$) (data not shown). Furthermore, the prevalence of HE phenotype distribution between EUB and AFB among A_1 blood donors was not significant (EUB, LE = 141, HE = 2; AFB, LE = 39, HE = 1; Fisher's exact test, $P = 0{\cdot}5251$).

Histogram patterns and comparison of antigen quantity on platelets by expression level

A antigen. Most of the LE platelet phenotype of A_1 subgroup (62·2%) showed very low MFI and peak channel (PC) profiles, chiefly in the negative region of the distribution, expressing an asymmetric distribution skewed to the right. However, the platelets (PLTs) of almost 30% of the donors showed a wide variation of MFI and PC profiles, ranging from low values near the negative region of the distribution to those lying borderline with the HE platelet phenotype values, denoting a bimodal histogram with high density of A antigen sites. The HE phenotype among A_1 donors has shown a distinctive separate and symmetric distribution with high MFI and PC profiles, expressing a relatively high density of A antigen sites (Fig. 2a–c).

Most of the platelets of the A_{int} individuals ($n = 11$) have shown profiles very similar to those observed among most of the LE platelet phenotypes of A_1 subgroup. However, there were a few individuals ($n = 6$) with a slightly higher density of A antigen sites (Fig. 2e,f). The platelets of group O and subgroup A_2 donors could not be distinguished by visual inspection of the histogram (data not shown).

H antigen. The histogram pattern for all the A variants were similar to a unimodal distribution, reflecting similar

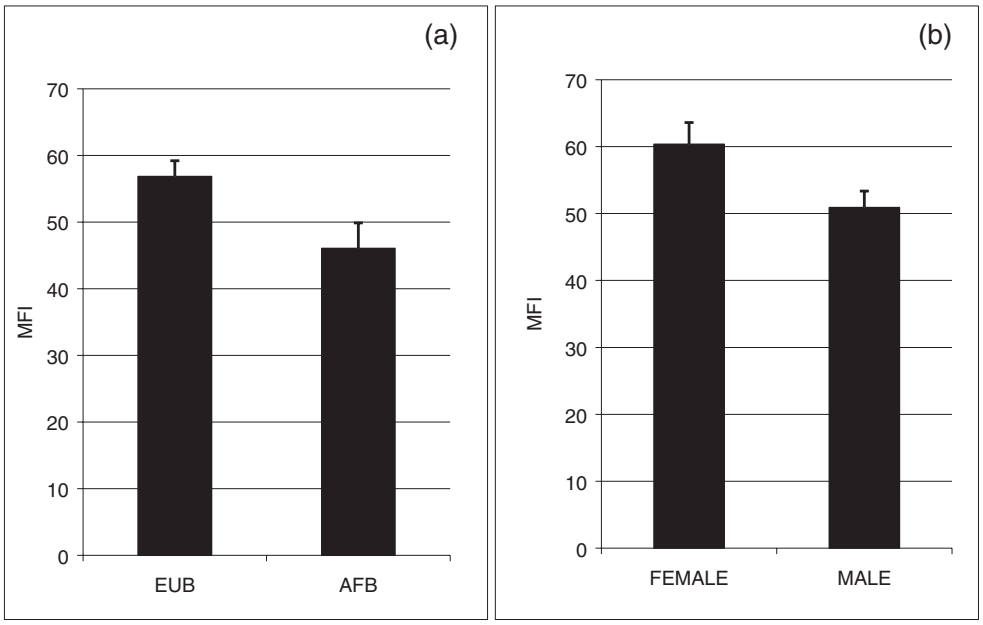


Fig. 1. Ethnic and sexual differences. (a) Ethnic differences. The amount (MFI) of A antigen on platelets of the A_1 subgroup between EUB ($n = 143$) and AFB ($n = 40$) was significantly different (Mann–Whitney = 2111·5; $Z = 2{\cdot}53$, $P = 0{\cdot}0115$). No significant differences were found for the A_2 subgroup (EUB, $n = 33$; AFB, $n = 8$) (Mann–Whitney = 108·0; $Z = 0{\cdot}79$, $P = 0{\cdot}4296$). A_{int} subgroup was not analysed due to the small sample size ($n = 17$). (b) Sexual differences. The amount (MFI) of A antigen expressed on platelets of the A_1 subgroup between male ($n = 117$) and female ($n = 66$) was significant (Mann–Whitney = 2867·5; $Z = 2{\cdot}89$, $P = 0{\cdot}0039$). Bars denote ±1 SEM.

and very high MFI and PC profiles (Fig. 2g–i). Both patterns (A and H antigens) were reproducible in blood samples drawn along the time.

Distribution of the A fluorescence intensity on platelets among subgroups

A_1, A_2 and A_{int} donors differ significantly among themselves (Fig. 3a). Differences between the MFI on platelets of A_1 and A_2 ($P < 0{\cdot}0001$), A_1 and A_{int} ($P = 0{\cdot}0011$), and A_2 and A_{int} ($P = 0{\cdot}0055$) donors were also significant. Furthermore, PLTs of A_2 donors differ significantly from those of O blood group individuals (Mann–Whitney A_2 vs O 357·0; $Z = 5{\cdot}33$, $P < 0{\cdot}0001$). The strength of this reactivity on RBC's control cells showed the same pattern as on platelets, however, with higher values (Fig. 3b).

Distribution of the H fluorescence intensity on platelets among subgroups

Several control tests were performed with RBCs to verify the standardisation of *U. europaeus* anti-H lectin. These studies confirmed the expected results with the working dilution of the anti-H lectin. The gradient of the H reactivity of RBCs shows the classical results expected for *U. europaeus* lectin (Fig. 3b). However, the H reactivity of the platelets does not follow a regular pattern and no significant differences were observed (Fig. 3a).

Distribution of A and H MFI within subgroups

There are two ways to analyse the relationship of A and H content of platelets, one of them was described above comparing the respective amount of intensity of fluorescence among subgroups (Fig. 3a,b). The other approach is to examine the correlation and linear regression of both the parameters. Under this approach, linear regression for both, EUB and AFB, is shown in Fig. 4.

Variability of MFI individual values

The expresser character may manifest the inherited tendency to remain the same (phenotypic stability) and the tendency to change in response to current environmental conditions. The test for correlated samples is useful in removing the effects of pre-existing individual differences (named SS_{wg} or subjects, extraneous to the principal question of the test), leaving only the error (random variability). The variability inside the samples reflects the fact that there is substantial difference

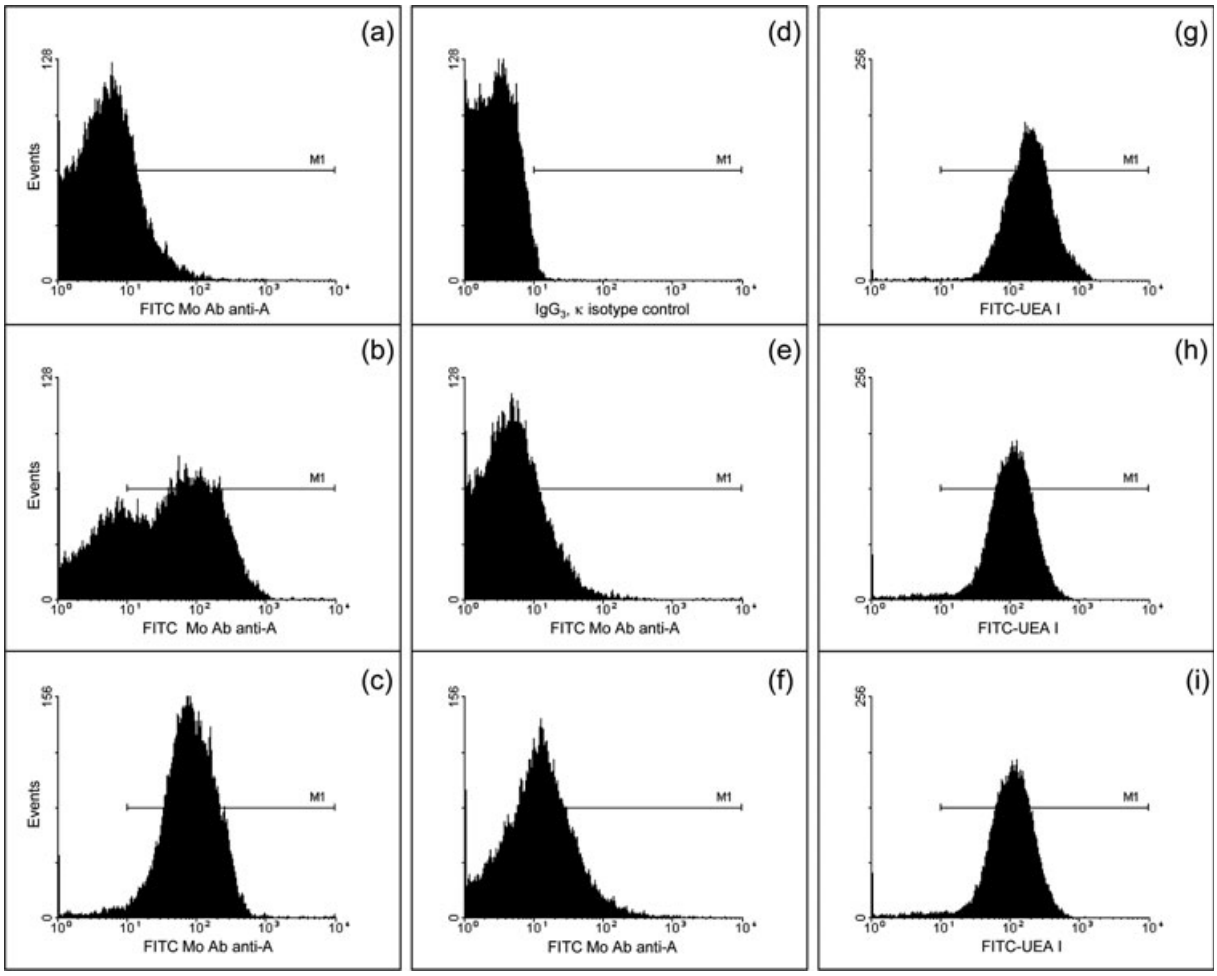


Fig. 2. Histogram patterns for the blood group A and H antigen expression on platelet membranes. Platelets were dual labelled with (i) a platelet-specific monoclonal antibody (PE-CD41) and FITC-labelled anti-human blood group A monoclonal antibody (a–c, e and f) and (ii) PE-CD41 and FITC anti-H *U. europaeus* lectin (FITC-UEA I) (g–i). (a and b) The variability patterns of the LE phenotype and (c) the HE phenotype for the A_1 subgroup. (d) IgG3κ isotype control for anti-human blood group A. (e and f) Sample histograms of platelets of A_{int} population. The last column (g–i) shows the same platelet samples shown in a–c exposed to the FITC-UEA I.

among people with respect to the variable fluorescence intensity. Samples of platelets from 20 individuals of A_1 subgroup were studied at 0, 30 and 45 days. The results showed that between groups SS = 0·2749, d.f. = 2, MS = 0·1375, $F = 2{\cdot}1756$, $P = 0{\cdot}127479$, where SS is the squared sum and MS is the mean square. The variability within days was great; however, when the pre-existing individual differences within each day were removed, the remaining value, the error, which represents the effect of the random variability, and the value of $F = \mathrm{MS}_{\mathrm{between}}/\mathrm{error} = 0{\cdot}1375/0{\cdot}0632 = 2{\cdot}1756$ were not significant. Days 0, 30 and 45 have shown a Gaussian distribution of the ln of the MFI individual values (data not shown). One-way ANOVA for three correlated samples also showed no significant variation ($P = 0{\cdot}127479$). This result indicated that there was stability of the MFI values, suggesting an inherited condition.

EM algorithm

The values for the MFI of A antigen on platelets of each blood group A_1 individual were subjected to a likelihood ratio test using the EM algorithm (Pawitan, 2001). The results of the likelihood ratio χ^2 test statistics when testing a single normal vs a two-component normal mixture [$\chi^2_{\mathrm{d.f.}=5-2=3} = -2(l_2 - l_1) = -2(-123{\cdot}0930 + 122{\cdot}1244) = 1{\cdot}9372$, $P = 0{\cdot}5 - 0{\cdot}6$] or a three-component normal mixture [$\chi^2_{\mathrm{d.f.}=8-2=6} = -2(l_1 - l_3) = -2(-123{\cdot}0930 + 118{\cdot}6598) = 8{\cdot}7944$; $P = 0{\cdot}1 - 0{\cdot}2$]

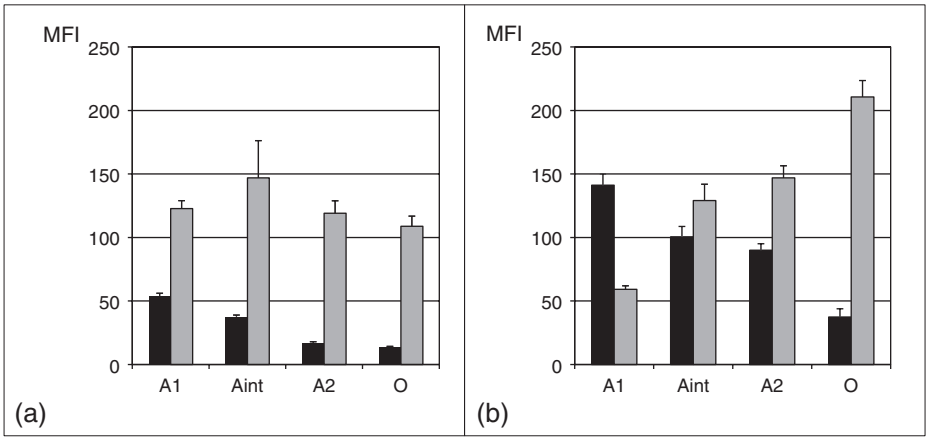


Fig. 3. Blood group A and H expression on platelets and RBCs. (a) On platelet surface. The fluorescence intensity on platelets differs significantly among subgroup A_1, A_2 and A_{int} donors (Kruskal–Wallis = 90·77, d.f. = 2, $P < 0·0001$) when stained with anti-human blood group A monoclonal antibody (black columns). A decreasing trend of A strength may be observed. The reactivity of *U. europaeus* lectin with platelets (grey columns) does not follow a regular pattern and no significant differences were observed when the H content of the A variants was compared with O blood group donors (Kruskal–Wallis = 1·73, d.f. = 3, $P = 0·6299$). (b) On RBC surface. The fluorescence intensity on RBCs differs significantly among subgroup A_1, A_2 and A_{int} donors (Kruskal–Wallis = 50·86, d.f. = 2, $P = 0·0001$) when stained with anti-human blood group A monoclonal antibody (black columns). A decreasing trend of A strength and an increasing trend of H (*U. europaeus*) (grey columns) may be observed. Significant differences were observed when the H content of the A variants was compared (Kruskal–Wallis = 44·69, d.f. = 2, $P < 0·0001$). Bars denote ±1 SEM.

were non-significant and also when comparing a two-component normal mixture and a three-component normal mixture [$\chi^2_{\text{d.f.}=8-5=3} = -2(l_2 - l_3) = -2(-122·1244 + 118·6958) = 6·8572$, $P = 0·05 - 0·1$], the result was non-significant. The graphic representation is shown in Fig. 5. It was estimated that on average, 98·4% of the ln MFI values, representing the LE platelet phenotype, were lower than or equal to the mean + 2SD, which is in agreement with the normal distribution theory.

Akaike's information criteria

When comparing a series of models specified *a priori*, the one with the lowest AIC is the 'best' one for the data at hand, being a statistical procedure that provides a measure of the goodness-of-fit of a model to a set of data (Burnham & Anderson, 2002). $AIC = -2(LL - p)$, where LL is the log likelihood at the maximum likelihood fit and p is the number of parameters of the model. It is noteworthy that the same data set was used for each model, i.e. the same observations were used for each analysis. For A_1 blood donors, the data for a single normal [$AIC = -2(-123·0930 - 2) = 250·186$] vs a two-component normal mixture [$AIC = -2(-122·1244 - 5) = 254·2488$] and vs a three-component normal mixture [$AIC = -2(-118·6958 - 8) = 253·3916$] have also shown that the 'best' model was that of one single normal distribution (Fig. 5). These results were similar to those obtained with the generalised likelihood ratio test.

AFB family results

A family study was performed on three generations of relatives of the HE group A_1 *propositus* who demonstrated the largest amount of platelet A antigen. Blood group A antigen levels on the RBCs of these family members were similar to those of five platelet LE AFB donors as measured by serological titrations (titre, $P = 0·6623$; titration score, $P = 0·8182$). However, the serum GTA from the three HE family members tested (II-3, II-5 and III-1) was capable of converting more H antigen on group O RBCs into A antigens compared with the control GTA from A_1 LE donors, as measured by both serological titrations ($\log_{10}$ end-point titre mean = 1·88 vs 1·69; $P = 0·0145$) and flow cytometry (MFI = 33·79 vs 17·93; $P = 0·0026$) (data not shown). All HE family members were secretors, based on either their Lewis phenotypes or their saliva inhibition studies (Fig. 6a).

The comparison between the A and H antigen content of the HE individuals from the population sample with those of the family members has shown that the value of A MFI of the population sample HE (mean = 152·93) falls outside the range of the mean of the HE of the

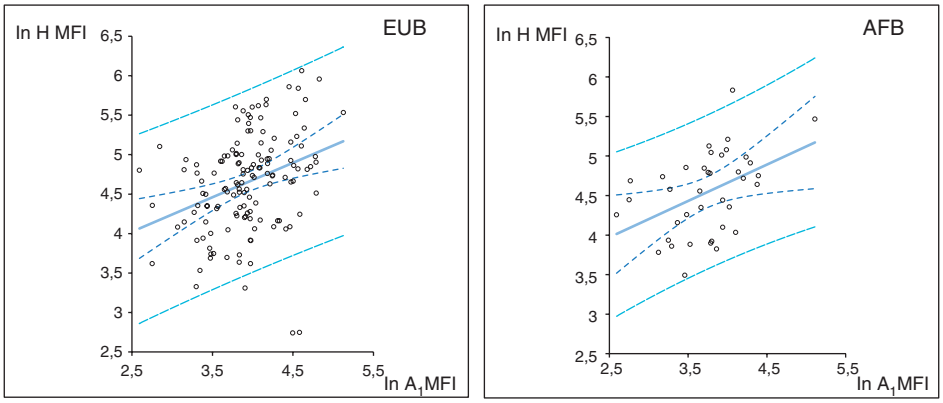


Fig. 4. Linear regression of blood group A and H antigen expression on platelets of the A_1 subgroup. The amount of A and H antigens expressed on platelets of the A_1 subgroup are positively correlated, i.e. when A antigen increases, H antigen also increases ($P < 0{\cdot}001$). Shown on the left is the A_1 blood group of the EUB population. Solid linear fit $= 2{\cdot}934 + 0{\cdot}436x$. Shown on the right is the A_1 blood group of the AFB population. Solid linear fit $= 2{\cdot}823 + 0{\cdot}4596x$. Dotted lines represent 95% CI; dashed lines represent 95% prediction interval. The other subgroups did not show correlation between A and H antigens expressed on platelets (A_{int}, $P = 0{\cdot}1407$; A_2, $P = 0{\cdot}6168$).

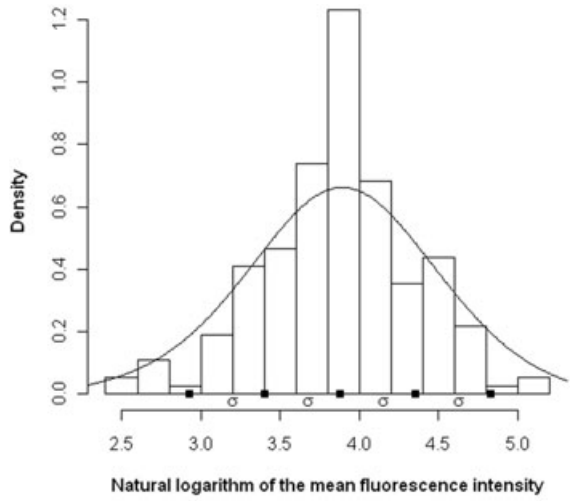


Fig. 5. Kernel-smoothed curve of A antigen expression on platelets from A_1 blood donors. A single normal model, with a mean of the $3{\cdot}88 \pm 0{\cdot}48$ (ln), fitted the data better than a mixture of two or three distributions when submitted to the likelihood ratio test by EM algorithm.

family components [mean = 336·97, 95% confidence interval (CI) 204·80–556·70]. This mean is about 6·4 times higher than the mean of the LE phenotype (52·46). In relation to the H content of the HE expressers, the HE of the family components shows a lower content than the population sample: the mean of the H content of the HE of the family ($30{\cdot}57 \pm 5{\cdot}32$) falls outside the 95% CI of the LE fraction of the population (mean = 119·10, 95% CI 109·46–128·44). The expression of the A and H antigens on both, platelets and RBCs, obtained with the *propositus* blood sample, are shown (Fig. 6b,c).

Genomic DNA was available for analysis from 8 of 14 of the living family members including 4 of 7 of the HE family members (I-1, II-5, III-5 and III-6), whose genotypes are indicated in Fig. 6a. By PCR-RFLP and PCR-ASP analyses, all four of these HE members were heterozygous for an A^1 allele with its single-unit enhancer region [*A101*, according to the other terminology used in the Blood Group Antigen Mutation Database (http://www.bioc.aecom.yu.edu/bgmut)]. Direct DNA sequencing of all seven exons and the flanking intronic regions of the *ABO* gene in two HE family members revealed a consensus A^1B (*A101/B101*) genotype in one member (I-1) and a consensus A^1O^1 genotype in the other (II-5); no unexpected polymorphisms were detected. In addition, the entire coding region of both *FUT1* and *FUT2* were also analysed in these two members and no unexpected polymorphisms were identified (Yazer *et al.*, 2005).

DISCUSSION

In some donors, our studies did not disclose clear-cut differences between the LE and HE-I phenotypes, as described by Curtis *et al.* (2000). The HE-I phenotypes from the point of view of MFI values and also from the histogram patterns cannot always be differentiated from the LE phenotype. For these reasons, we have included both of them as an LE phenotype. On account of these experimental results, we included in the LE platelet phenotype a large spectrum of MFI values.

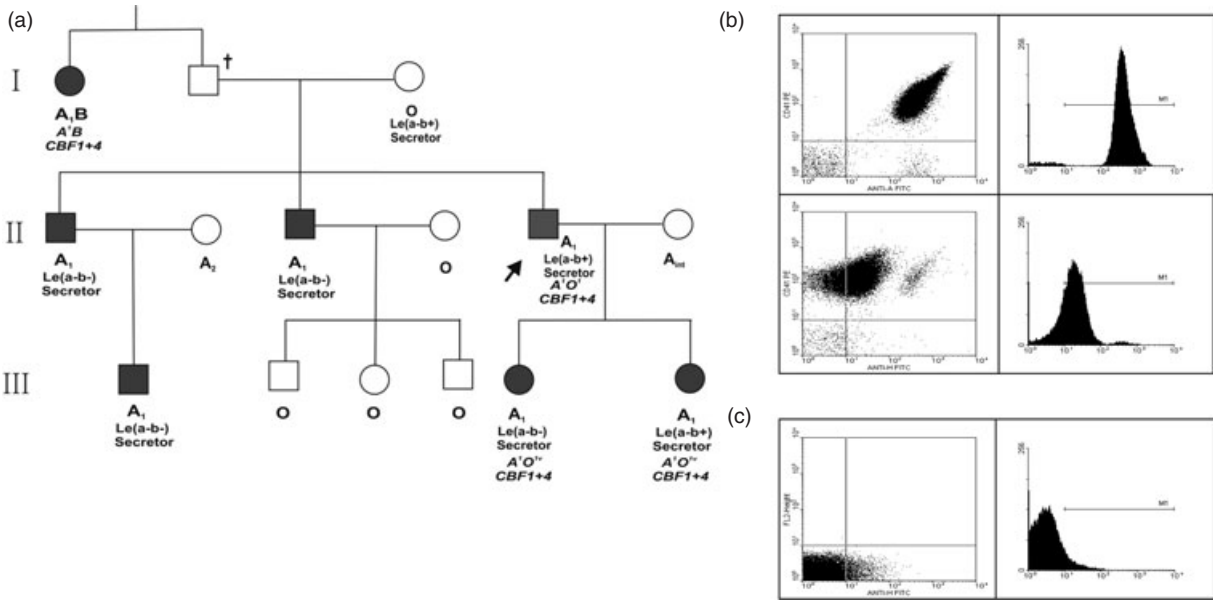


Fig. 6. Pedigree of the AFB family and A and H blood group antigen expression in the *propositus* blood sample. (a) Pedigree of the family. The shaded boxes indicate the HE individuals determined by flow cytometry. The phenotypes are presented in normal font, whereas genotypes are presented in italics and the arrow indicates the *propositus*. CBF refers to the upstream enhancer region of the *ABO* gene, and the numbers 1 and 4 reflect the number of sequences within this region, i.e. 1 + 4 would indicate a heterozygote for both a single enhancer unit and a four-repeat enhancer region (see text). (b) Platelets of the *propositus*. Platelets were dual labelled with (i) a platelet-specific monoclonal antibody (PE-CD41) and FITC-labelled anti-human blood group A monoclonal antibody (on the top) and (ii) PE-CD41 and FITC anti-H *U. europaeus* lectin (FITC-UEA I) (on the bottom). (c) RBCs of the *propositus*. RBCs, when exposed to the FITC-UEA I. All HE individuals in this family demonstrated flow cytometry histograms similar to the A_1 HE *propositus*.

The A reactivity on platelets follows the same gradient of MFI values as the respective RBC's control cells, which was also described by Cooling (2006).

A positive correlation was observed between A_1 and H variables (Fig. 4). Our results are similar to those reported by Cooling *et al*. (2005). The extremely low levels of the H content on platelets of the family HE members (in relation to the amount observed in the population sample) (Fig. 6b, on the bottom – *propositus*/ Fig. 2g–i – population sample), together with the almost virtual absence of H antigen on RBCs (Fig. 6c) and the higher levels of serum α-GTA, all agree with the HE-II platelet phenotype described by Curtis *et al*. (2000, 2008). Findings made in this family have shown that these platelets carry significantly higher levels of A antigen than other A individuals and have indicated that essentially all H substance, the substrate for synthesis of A antigen, was converted to A by the higher active α-GTA.

Furthermore, our results – A_2 and A_{int} platelet donors have MFI means significantly different from those obtained with platelets from group O individuals – are at odds with other reports which have indicated that the amount of A antigen on most platelets from A_2 subgroup was very weak or undetectable (Holgersson *et al*., 1990; Ogasawara *et al*., 1993; Hou *et al*., 1996; Curtis *et al*., 2000; Julmy *et al*., 2003). An explanation of this difference is apparently not related to ethnic origins. Furthermore, the A strength on platelets suggests that, at least in our population, it may not be appropriate to transfuse platelets from either A_2 or A_{int} blood donors to group O patients (chiefly, if they receive multiple transfusions for a rather long time), unless these patients have a previous test showing weak or null anti-A agglutinin titration scores.

What remains to be elucidated is the mechanism of inheritance and the factor(s) responsible for producing the HE platelet phenotype. The occurrence of HE in several members suggests familial aggregation. Furthermore, in repeated measures, stability of the MFI values suggesting an inherited condition was observed. Polygenic inheritance – which cannot be excluded as a hypothesis – is demonstrated in the transmission pattern of other continuous or quantitative traits, such as blood pressure (Lander & Schork, 1994). However, the inheritance of the trait was considered to be transmitted as a dominant Mendelian character related to the A or B antigen (Ogasawara *et al*., 1993; Curtis *et al*.,

2000). We believe that since the first descriptions of the trait (Ogasawara *et al*., 1993; Curtis *et al*., 2000; Cooling *et al*., 2005), the intensity of fluorescence, a typical continuous variable, was described as a practical approximation to a discrete form of two or three phenotypes. The expresser phenotype appears to be cosegregating with the A^1 and *B* alleles in the families described by others (Ogasawara *et al*., 1993; Curtis *et al*., 2000, 2008) and in this study. However, this cosegregation does not necessarily mean that the A^1 and *B* alleles are conditioning the expresser phenotypes. Against the hypothesis of a trait linked to the *ABO* locus, there is the apparent absence of recombinants in the A or B families.

The pattern of inheritance of the HE platelet phenotype demonstrated in the family study (Fig. 6a) suggests autosomal dominant inheritance of the HE phenotype-producing factor(s). Similar pedigrees in Japanese donors demonstrating an inheritance pattern consistent with autosomal dominance have also been reported (Ogasawara *et al*., 1993). This proposed mode of inheritance whereby the HE-producing factors affect either the GTA or GTB individually explains why, of the 106 AB Japanese donors studied, none demonstrated simultaneously high expression of both A and B antigens on their platelets (Ogasawara *et al*., 1993), a fact that argues for an *ABO* locus-linked inheritance as opposed to an independent factor able to 'boost' the action of any ABO transferase present.

Furthermore, Cooling *et al*. (2005) and the present report found a linear relationship between the amount of A and H antigens on the platelet surface of group A_1 donors. We hypothesised that a hyperactive 2-α-L-fucosyltransferase encoded by the *FUT1* or *FUT2* loci might be responsible for producing the HE phenotype by providing larger quantities of H antigen in secretions and on RBCs, respectively, which could then be converted into group A antigens. However, complete analysis of the *FUT1* and *FUT2* coding regions in two HE individuals revealed only genes without any unexpected polymorphisms consistent with normal H synthesis and thus do not appear to influence the expresser status. All the HE members of the family studied were secretors of A and H substances, thus the passive adsorption of A antigens onto the platelets (Dunstan *et al*., 1985; Kunicki & George, 1989) cannot be definitively excluded as a contributing factor of the HE state, but it must be realised that based on the frequency of secretor status, this could merely be a coincidence in this family.

The current data set refutes several previously proposed mechanisms of producing the HE phenotype and points to other theories of its inheritance. There are an allele-specific number of minisatellite repeats in the 5′ enhancer region of the *ABO* allele (Kominato *et al*., 1997; Irshaid *et al*., 1999). Those enhancers featuring four repeats are associated with higher rates of transcription in luciferase reporter gene experiments (Yu *et al*., 2000). With this in mind, we decided to analyse the 5′ promoter region in the HE family members for which DNA was available. In an earlier study, which was reported only in abstract form, the presence of the theoretically favourable four-repeat variable number of tandem repeats (VNTR) enhancer region associated with an A^1 allele was suggested to cause HE phenotype in two donors (Gershan *et al*., 2006). To date, this finding has not been presented in full or confirmed by other investigators. Irshaid *et al*. (1999) screened 234 donors of different ethnic/geographic origins and found only a single Jordanian donor who had an A^1 allele in *cis* with four repeats. Since then, at least twice as many have been screened and yet no four-repeat enhancers have been found associated with the A^1 allele (M. L. Olsson, personal communication). This does not appear to parallel the suggested frequency of HE phenotypes (4–7% in different studies) if the A^1 allele with four repeats really had a role in the molecular basis of HE. It should also be noted that there are currently no experimental data to support the notion that enhancer repeats are important for *ABO* transcription in haematopoietic tissue. The studies suggesting that they do in fact play a role in regulating the rate of transcription (and thereby theoretically, e.g. the A or B expression on platelets) were performed in cell lines of gastric cancer origin (Kominato *et al*., 1997; Yu *et al*., 2000).

In this study, the expected single-unit enhancer region associated with the A^1 allele of HE individuals, the exons 1–7, including splice sites and the putative proximal promoter region had no unexpected mutations/polymorphisms. On the contrary, a closer analysis of the current data reveals some interesting phenomena that might not necessarily be accounted for by a simple inheritance pattern. In spite of the consensus *ABO* gene in the current platelet HE family members (I-1 and II-5 individuals), the serum GTA from other HE family members (II-3 and III-1 individuals) synthesised more A antigens on group O cells compared with the serum GTA from A_1 LE control donors. The higher than normal GTA activity has also been previously reported in other platelet HE individuals (Ogasawara *et al*., 1993; Curtis *et al*., 2000), but it is still unclear if this finding is due to a higher concentration of serum GTA or a qualitatively enhanced enzyme at the same concentration as in LE individuals. Yet another possibility to explain the varying ABO antigen levels on platelets is megakaryocyte-related interindividual differences.

The cause of the platelet phenotype remains elusive, although its aetiology probably lies outside the *ABO* locus itself. We have presented data that support both a single and polygenetic explanation for this phenotype;

further research is warranted to more fully elucidate the cause of this fascinating and potentially clinically relevant phenomenon.

ACKNOWLEDGMENTS

We thank the members of the family who kindly provided information, blood and saliva samples. We also thank Drs Åsa Hellberg, Martin L. Olsson and Jill Storry from the Division of Hematology and Transfusion Medicine, Faculty of Medicine, Lund University & Blood Centre, University Hospital, Lund, Sweden, and Dr Mark Yazer from the Institute for Transfusion Medicine and Department of Pathology, University of Pittsburgh, Pittsburgh, USA, for performing the *ABO* and *FUT-1* genetic experiments.

This research was supported by Conselho Nacional de Pesquisa e Desenvolvimento Tecnológico (CNPq), Fundação Coordenação para o Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasília and Fundação Universitária José Bonifácio (FUJB), Rio de Janeiro.

REFERENCES

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer-Verlag, New York.

Cooling, L.L.W., Kelly, K., Barton, J., Hwang, D., Koerner, T.A.W. & Olson, J.D. (2005) Determinants of ABH expression on human blood platelets. *Blood*, **105**, 3356–3364.

Cooling, L.L. (2006) ABH expression on human blood platelets [letter]. *Blood*, **107**, 842.

Curtis, B.R., Edwards, J.T., Hessner, M.J., Klein, J.P. & Aster, R.H. (2000) Blood group A and B antigens are strongly expressed on platelets of some individuals. *Blood*, **96**, 1574–1581.

Curtis, B.R. & Aster, R.H. (2006) Expression of ABH antigens on platelets [letter]. *Blood*, **107**, 842–843.

Curtis, B.R., Fick, A., Lochowitcz, A.J., McFarland, J.G., Ball, R.H., Peterson, J. & Aster, R.H. (2008) Neonatal alloimmune thrombocytopenia associated with maternal–fetal incompatibility for blood group B. *Transfusion*, **48**, 358–364.

Daniels, G. (2002) *Human Blood Groups.* Blackwell Publishing Co., Oxford.

Dunstan, R.A., Simpson, M.B., Knowles, R.W. & Rosse, W.F. (1985) The origin of ABH antigens on human platelets. *Blood*, **65**, 615–619.

Gershan, J.A., Visentin, G.P., Curtis, B.R. & Aster, R.H. (2006) Molecular basis for high expression of A_1 antigen on platelets of some normal individuals [abstract]. *Blood*, **98**, 708A–709A.

Holgersson, J., Breimer, M.E., Jacobsson, A., Svensson, L., Ulfvin, A. & Samuelsson, B.E. (1990) Glycolipid- and glycoprotein-based A antigen expression in human thrombocytes: A_1/A_2 difference. *Glycoconjugate Journal*, **7**, 601–608.

Hou, M., Stockelberg, L., Rydberg, J., Kutti, J. & Wadenvik, H. (1996) Blood group A antigen expression in platelets is prominently associated with glycoprotein Ib and IIb: evidence for an A_1/A_2 difference. *Transfusion Medicine*, **6**, 51–59.

Irshaid, N.M., Chester, M.A. & Olsson, M.L. (1999) Allele-related variation in minisatellite repeats involved in the transcription of the blood group *ABO* gene. *Transfusion Medicine*, **9**, 219–226.

Julmy, F., Achermann, F., Schulzki, T., Carrel, T. & Nydegger, U. (2003) PLTs of blood group A_1 donors express increased surface A antigen owing to apheresis and prolonged storage. *Transfusion*, **43**, 1378–1385.

Kominato, Y., Tsuchiya, T., Hata, N., Takizawa, H. & Yamamoto, F. (1997) Transcription of human *ABO* histo-blood group genes is dependent upon binding of transcription factor CBF/NF-Y to minisatellite sequence. *The Journal of Biological Chemistry*, **272**, 25890–25898.

Kunicki, T.J. & George, J.N. (1989) *Platelet Immunobiology: Molecular and Clinical Aspects.* JB Lippincott, Philadelphia, PA.

Lander, E.S. & Schork, N.J. (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048. Erratum in: *Science*, **266**, 353.

Mallory, D. (Ed-in-Chief) (1993) *Immunohematology: Methods and Procedures.* American Red Cross, Rockville, MD.

Metcalfe, P., Williamson, L.M., Reutelingsperger, C.P.M., Swann, I., Ouwehand, W.H. & Goodall, A.H. (1997) Activation during preparation of therapeutic platelets affects deterioration during storage: a comparative flow cytometric study of different production methods. *British Journal of Haematology*, **98**, 86–95.

Michelson, A.D., Barnard, M.R., Krueger, L.A., Frelinger, A.L. & Furman, M.I. (2000) Evaluation of platelet function by flow cytometry. *Methods*, **21**, 259–270.

Michelson, A.D. (2006) Evaluation of platelet function by flow cytometry. *Journal of Pathophysiology of Haemostasis and Thrombosis*, **35**, 67–82.

Mody, M., Lazarus, A.H., Semple, J.W. & Freedman, J. (1999) Preanalytical requirements for flow cytometric evaluation of platelet activation: choice of anticoagulant. *Transfusion Medicine*, **9**, 147–154.

Ogasawara, K., Ueli, J., Takenaka, M. & Furihata, K. (1993) Study of the expression of ABH antigens on platelets. *Blood*, **82**, 993–999.

Olsson, M.L. & Chester, M.A. (1995) A rapid and simple *ABO* genotype screening method using a novel B/O^2 versus A/O^1 discriminating nucleotide substitution at the *ABO* locus. *Vox Sanguinis*, **69**, 242–247.

Olsson, M.L. & Chester, M.A. (1996) Frequent occurrence of a variant O^1 gene at the blood group *ABO* locus. *Vox Sanguinis*, **70**, 26–30.

Olsson, M.L., Hosseini-Maaf, B., Hellberg, Å. & Chester, M.A. (1998) Potential genotyping errors caused by recombinant hybrid alleles at the *ABO* locus [abstract]. *Transfusion*, **38**, 3S.

Olsson, M.L., Irshaid, N.M., Hosseini-Maaf, B., Hellberg, Å., Moulds, M.K., Sareneva, H. & Chester, M.A. (2001)

Genomic analysis of clinical samples with serological ABO blood grouping discrepancies: identification of 15 novel *A* and *B* subgroup alleles. *Blood*, **98**, 1585–1593.

Palatnik, M. (1984) A and AB, ABO blood group variants in Brazil. *Brazilian Journal of Genetics*, **7**, 727–733.

Palatnik, M., Silva Junior, W.A. da Estalote, A.C., de Oliveira, J.E., Milech, A. & Zago, M.A. (2002) Ethnicity and type 2 diabetes in Rio de Janeiro, Brazil, with a review of the prevalence of the disease in Amerindians. *Human Biology*, **74**, 533–544.

Pawitan. Y. (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press, New York.

Skogen, B., Rossebø-Hansen, B., Husebekk, A., Havnes, T. & Hannestad, K. (1988) Minimal expression of blood group A antigen on thrombocytes from A_2 individuals. *Transfusion*, **28**, 456–459.

Storry, J.R., Samuelsson, J., Olsson, M.L., Rodrigues, M.J., Levene, C., Yahalom, V., Strindberg, J., Bullock, T. & Poole, J. (2004) Heterogeneity at the *FUT1* and *FUT2* loci: genetic basis of three ethnically diverse H-deficient individuals [abstract]. *Transfusion*, **44**(Suppl.), 26A–27A.

Yabe, R., Bannai, M., Nakata, K., Seno, T., Okubo, Y. & Yamaguchi, H. (1989) Family study and frequency of blood group with strong B transferase accompanied by decreased A and H antigens. *Blut*, **59**, 157–161.

Yazer, M.H., Sant'Anna Gomes, B.M., Hellberg, Å., Estalote, A.C., Palatnik, M., Pimenta, G., Pereira, B. de B., Chester, M.A. & Olsson, M.L. (2005) Investigation of *ABO* allele sequence enhancer, and gene zygosity in donors with high A antigen expression on platelets. *Transfusion*, **45**(Suppl.), 130A.

Yip, S.P., Chee, K.Y., Chan, P.Y., Chow, E.Y. & Wong, H.F. (2002) Molecular genetic analysis of para-Bombay phenotypes in Chinese: a novel non-functional *FUT1* allele is identified. *Vox Sanguinis*, **83**, 258–262.

Yu, L.C., Chang, C.Y., Twu, Y.C. & Lin, M. (2000) Human histo-blood group *ABO* glycosyltransferase genes: different enhancer structures with different transcriptional activities. *Biochemical Biophysical Research Communication*, **273**, 459–466.

Apêndice E

Redes Neurais Artificiais

Este Apêndice apresenta uma aplicação das redes neurais do tipo *feedforward* (Seção 2.5.1), para investigar a associação entre a poluição atmosférica e condições climáticas no número de internações hospitalares, por motivo de bronquiolite infantil:

NASCIMENTO, E. M., PEREIRA, B. DE B., SEIXAS, J. M. Redes Neurais Artificiais: Uma Aplicação no Estudo da Poluição Atmosférica e Seus Efeitos Adversos à Saúde. *Revista Brasileira de Biometria*, v. 27, pp. 37-50, 2009.

REDES NEURAIS ARTIFICIAIS: UMA APLICAÇÃO NO ESTUDO DA POLUIÇÃO ATMOSFÉRICA E SEUS EFEITOS ADVERSOS À SAÚDE

Emília Matos do NASCIMENTO[1]
Basilio de Bragança PEREIRA[1,2]
José Manoel de SEIXAS[3]

- RESUMO: Há uma enorme necessidade em buscar a associação entre condições climáticas e poluição atmosférica com mortalidade ou internações hospitalares por doenças respiratórias. Este trabalho propõe a utilização das redes neurais como metodologia alternativa para avaliar tal associação. Os dados utilizados referem-se ao número de internações hospitalares na cidade de Paris, por bronquiolite infantil, entre 1997 e 2000. Os modelos neurais foram avaliados em termos de descrição dos dados e da sua capacidade de generalização. Os melhores resultados foram obtidos através do pré-processamento dos dados, com remoção de ciclos e uso de filtro de média móvel. Um estudo de relevância das variáveis explicativas foi também desenvolvido. Os resultados obtidos foram compatíveis aos encontrados através dos modelos aditivos generalizados, apontando o material particulado (PM_{10}) como principal responsável no número de internações hospitalares.
- PALAVRAS-CHAVE: Redes neurais artificiais; poluição atmosférica; doenças respiratórias.

1 Introdução

A cada dia aumenta a preocupação do homem com a degradação ambiental, que se expressa através de mudanças climáticas, poluição do ar, contaminação das águas e do solo, com conseqüências desastrosas para a fauna e a flora, causando doenças e afetando negativamente a qualidade da vida humana (Bates et al., 2008).

A poluição do ar tem sido apontada como um dos principais responsáveis por doenças relacionadas ao sistema respiratório, especialmente em crianças, idosos e pessoas com problemas respiratórios. A exposição aos poluentes pode provocar ou agravar doenças tais como asma, bronquite crônica, enfisema pulmonar, infecções pulmonares, rinofaringites e irritação nas vias respiratórias, entre outras, conduzindo a um elevado número de internações hospitalares, óbitos e aumentando a procura de atendimento nas salas de emergência (Singer et al., 2002; Šrám et al., 2005).

[1] Programa de Engenharia de Produção, COPPE/UFRJ, Caixa Postal: 68507, CEP. 21941-972 - Rio de Janeiro - RJ – Brasil. E-mail: *emilia@pep.ufrj.br*
[2] Faculdade de Medicina, HUCFF/UFRJ, Universidade Federal do Rio de Janeiro – UFRJ, Caixa Postal 68507, CEP: 21941-972, Rio de Janeiro, RJ, Brasil. E-mail: *basílio@hucff.ufrj.br*
[3] Laboratório de Processamento de Sinais, Programa de Engenharia Elétrica, COPPE/Poli-UFRJ, Caixa Postal: 68504, CEP. 21941-972, Rio de Janeiro, RJ, Brasil. E-mail: *seixas@lps.ufrj.br*

Vários episódios têm sido registrados em grandes centros urbanos em virtude das altas concentrações de poluentes na atmosfera, com trágicas conseqüências para a população (Braga et al., 2002; Conceição et al., 2002). A preocupação causada pelos efeitos decorrentes da degradação do ar tem sido motivo de discussão entre líderes de diversos países, levando-os a firmarem acordos para controle e redução da emissão de gases de efeito estufa (UNFCCC, 1992; UNFCCC, 2007).

A poluição atmosférica é um problema de saúde pública, constituindo um grande desafio à gestão. A compreensão dos seus efeitos pode contribuir para o planejamento e nortear a adoção de medidas públicas visando proteger a população. O conhecimento científico é de grande utilidade no auxílio à gestão da informação relativa à saúde coletiva, subsidiando a elaboração de medidas voltadas à prevenção e à redução da poluição do ar.

Este trabalho tem como objetivo propor a utilização das redes neurais artificiais (Haykin, 2008; Pereira e Rodrigues, 1998), como metodologia alternativa, no estudo da poluição atmosférica e seus efeitos adversos à saúde. As redes neurais artificiais foram utilizadas, para reproduzir a análise de Willems et al. (2007), que utilizaram os modelos aditivos generalizados (Hastie e Tibshirani, 1990) para estimar os efeitos causados pela poluição atmosférica e condições climáticas, tendo em vista o número de internações hospitalares na cidade de Paris, França, motivadas por bronquiolite infantil, doença respiratória causada frequentemente pelo vírus *syncethial respiratory virus* (RSV). O contato com este vírus causa, normalmente, um resfriado. Mas em crianças e, em algumas circunstâncias, especialmente no início do inverno, o vírus pode ser responsável por uma grave doença respiratória, conduzindo a um elevado número de consultas e internações hospitalares (Everard et al., 1994; Farhat et al., 2002; Gutiérrez et al., 2003).

A base de informação utilizada no presente estudo é apresentada na seção 2. A etapa de normalização e o pré-processamento dos dados para o projeto da rede neural são descritos na seção 3. A modelagem neural e a metodologia utilizada para identificação das variáveis relevantes são apresentadas na seção 4. Finalmente, na seção 5, são apresentados os resultados obtidos e, em seguida, as respectivas conclusões.

2 Base de dados

Este estudo foi desenvolvido sobre a mesma base de dados utilizada por WILLEMS et al. (2007), obtida no ERBUS (*Epidémiologie et Recueil des Bronchiolites en Urgence pour Surveillance*), referente a 43 hospitais localizados em Paris, no período compreendido entre 15 de outubro a 15 de janeiro dos anos de 1997 a 2000, além de dados meteorológicos e de poluição atmosférica medidos pelas estações de AIRPARIF - *Surveillance de la Qualité de l'Air en Ile-de-France*, nos anos de 1997 a 2001. O procedimento, adotado pelos referidos autores para evitar problemas com dados faltantes, foi utilizado também neste trabalho. Assim, foram considerados, ao todo, 419 dias de observação dos 34 hospitais, cujos dados encontravam-se completos. A base de informação é composta por 17 variáveis explicativas (12 variáveis climáticas e 5 de poluição atmosférica). A série histórica do número de internações hospitalares, motivadas por bronquiolite infantil, representa o alvo (variável dependente). A Figura 1 mostra essa série, indicando claramente a existência de variações sazonais. Quatro períodos podem ser observados, cada um dos quais representa um período incluindo o inverno europeu.

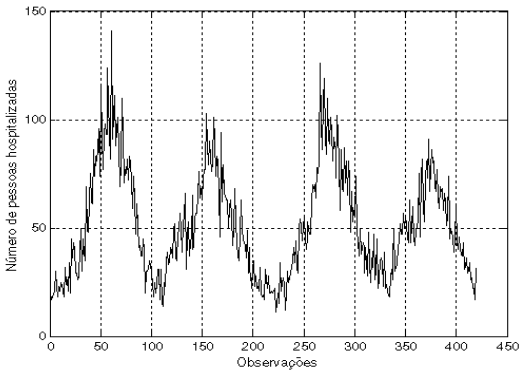


Figura 1 - Número de internações hospitalares.

3 Normalização e pré-processamento dos dados

Os modelos neurais foram projetados com a finalidade de se obter uma descrição dos dados ou uma generalização para o conjunto total dos dados. Nos modelos projetados para descrição, a avaliação foi feita em termos de interpretação dos dados, tendo sido utilizada toda a informação disponível no seu desenvolvimento. Os modelos que visam à generalização são restritos a um conjunto de desenvolvimento (treino), sendo a avaliação da generalização do modelo obtida através do teste de desempenho no conjunto restante de dados. Assim, nestes modelos, a base de dados foi subdividida em dois conjuntos: treinamento e teste. O conjunto de teste recebeu um terço das observações, tomadas de três em três, a partir da terceira, num total de 139 observações, cabendo ao conjunto de treinamento as 280 observações restantes, referentes aos índices 1, 2, 4, 5 da base de dados e assim, por diante.

O projeto das redes neurais artificiais consistiu em uma etapa inicial de normalização, que tem como finalidade adaptar os dados de entrada à faixa dinâmica das funções de ativação da rede neural, neste caso, tangente hiperbólica (na camada intermediária) e função linear (na camada de saída). O processo de normalização dos dados foi feito da seguinte forma:

$$x^* = \frac{x_0}{\bar{x} + n_s s} \tag{1}$$

onde x^* é a variável normalizada; x_0 é a variável original; $\bar{x}$ é a média obtida a partir da amostra que forma o conjunto de treinamento; s é o desvio-padrão da variável original, também obtido a partir do conjunto de treinamento; e n_s representa o número de desvios-padrão a considerar. O valor de n_s foi obtido empiricamente, respeitando a faixa dinâmica da rede neural, que deve conter o resultado encontrado no cálculo de x^*.

A normalização foi aplicada, inicialmente, ao conjunto de treinamento e, em seguida, estendida ao conjunto de teste, onde foi adotado o mesmo fator de normalização.

A característica sazonal das variáveis deve ser levada em consideração. Após a normalização, os dados devem ser dessazonalizados a fim de revelar o resíduo da série a ser modelada, permitindo assim, a detecção das características escondidas. Nelson et al. (1999) observaram, através da utilização de um modelo neural para previsão, que os resultados obtidos a partir de dados dessazonalizados são, significativamente, mais acurados do que os obtidos sem o pré-processamento. Calôba et al. (2002), Alekseev e Seixas (2008) procederam a dessazonalização dos dados através da remoção da tendência, seguida da retirada dos ciclos.

A tendência da série pode ser estimada pelo método dos mínimos quadrados, através do seguinte cálculo:

$$s_2[n] = s_1[n] - (a + b * n) \quad (2)$$

onde $s_1[n]$ é a série original normalizada; $s_2[n]$ é a série obtida após a eliminação da tendência; a é o intercepto; e b é o coeficiente angular da reta de regressão. Observando que a série alvo (número de internações hospitalares) não apresenta tendência (tal fato pode ser visualizado através da Figura 1), considerou-se desnecessária a retirada da tendência.

Os eventuais efeitos das concentrações dos poluentes e de possíveis variáveis de confundimento não incidem, necessariamente, no mesmo dia em que foi observado o evento (internação ou óbito). Assim, é também comum o uso de modelos com defasagem ou médias móveis das variáveis meteorológicas e dos poluentes (Zanobetti et al., 2000 apud Lima et al., 2001). Willems et al. (2007) observaram que a utilização de um filtro de médias móveis de seis dias forneceu um efeito mais significativo em seu modelo do que o uso de um filtro de médias móveis de dois dias, por exemplo. O filtro de médias móveis foi utilizado, também, neste trabalho.

Levando-se em consideração a característica sazonal das variáveis, avaliou-se a necessidade do pré-processamento dos dados. Esta avaliação foi feita após a implementação de diversas redes neurais, tendo, como critério, o maior poder de generalização do modelo. A generalização está associada à capacidade de predição de novos casos, indicando se o modelo neural conseguiu extrair as características principais da informação.

Os melhores resultados foram obtidos por um modelo no qual o pré-processamento dos dados foi feito através da retirada dos ciclos, utilizando a análise de Fourier, seguida do uso de um filtro de média móvel de seis dias.

A análise de Fourier (Cooley e Tukey, 1965) permite que sejam modelados os componentes que refletem os ciclos presentes na série. A retirada dos ciclos identificados pela análise de Fourier possibilita que a rede neural seja alimentada com a série residual, livre dos componentes já modelados. A remoção dos componentes senoidais de uma série é feita mediante a aplicação da transformada de Fourier e posterior observação das frequências que apresentem um pico significativo de amplitude, relativamente às demais frequências que compõem o espectro da série. Desta forma, identificados os coeficientes dos senos e co-senos que compõem a informação da frequência em questão, procede-se à retirada do pico correspondente.

Alekseev e Seixas (2008) removeram os ciclos observados na variável alvo do conjunto de treinamento, e estenderam esse procedimento às variáveis de entrada, quando a mesma frequência foi identificada. Calôba et al. (2002) utilizaram o valor obtido no

cálculo da raiz quadrada da média dos quadrados dos resíduos (RMSE) para decidir se um componente senoidal deveria, ou não, ser retirado. Desta forma, caso a remoção do ciclo provocasse um aumento no valor da RMSE, o componente senoidal não seria retirado.

Neste trabalho, os ciclos identificados na série alvo de treino foram retirados, deterministicamente, por senóides. Os mesmos ciclos, porventura presentes nas variáveis explicativas do conjunto de treinamento, foram também removidos. Após a retirada dos componentes senoidais das variáveis de entrada, uma nova análise espectral foi feita para verificar a existência de ciclos remanescentes. Caso tais ciclos fossem encontrados, seriam também removidos. De forma análoga ao procedimento adotado por Calôba et al. (2002), a decisão pela remoção, ou não, de um determinado componente senoidal ficou condicionada ao valor da RMSE. Assim, quando o aumento da RMSE foi observado, o componente senoidal não foi retirado. Esse procedimento foi estendido aos dados pertencentes ao conjunto de teste, retirando-se os mesmos componentes identificados no conjunto de treinamento.

4 Modelagem neural e análise de relevância

A implementação dos modelos foi feita através de redes neurais completamente conectadas, modelo MLP – *Multilayer Perceptron* (Rumelhart et al., 1986), com uma camada intermediária e sem realimentação. Os neurônios na camada escondida foram do tipo tangente hiperbólica. Na camada de saída, foi utilizado um único neurônio, com função de ativação linear. Os pesos da rede neural podem ser atualizados de duas formas: (1) modo batelada, onde a atualização dos pesos se dá quando todos os pares (entrada e saída) do conjunto de treinamento forem apresentados; e (2) modo instantâneo, onde a atualização ocorre cada vez que uma amostra do conjunto de treinamento for apresentada. Neste trabalho, o treinamento foi feito, no modo batelada, através do algoritmo *Resilient Backpropagation*, desenvolvido por Riedmiller e Braun (1993) e Riedmiller (1994), que tem a vantagem de convergir mais rapidamente.

A Figura 2 mostra a arquitetura do modelo neural, composto de 17 variáveis de entrada (12 relacionadas aos fatores climáticos e 5 referentes às concentrações de poluentes) e uma variável de saída (número de internações hospitalares).

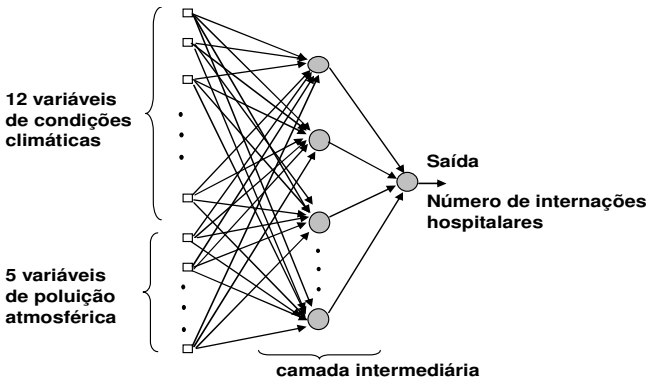


Figura 2 - Arquitetura do modelo neural.

Como medida de desempenho no processo de treinamento da rede neural, adotou-se o MSE (erro médio quadrático) cometido à saída da rede, em relação ao valor observado no alvo de treinamento. A função objetivo teve, como finalidade, a minimização do MSE no conjunto de treinamento.

O conjunto de teste não participa da fase de treinamento da rede neural, sendo usado para avaliar o poder de generalização da rede neural. O MAPE (erro percentual absoluto médio), calculado no conjunto de teste, foi adotado, tanto como critério de parada do treinamento, quanto para avaliar o poder de generalização do modelo. Esta medida é definida como:

$$mape = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (3)$$

onde y é o valor observado (variável resposta) e $\hat{y}$ é a saída da rede.

Após o treinamento do modelo neural, procedeu-se à recomposição dos alvos (número de internações hospitalares), adicionando-se, à saída obtida, os valores retirados na fase do pré-processamento (remoção de ciclos e sazonalidades) e multiplicando-se o resultado obtido pelo fator de normalização. O poder de generalização da rede foi medido no conjunto de teste através do cálculo do MAPE, avaliado sobre os alvos originais e os alvos recompostos.

A seleção das variáveis explicativas a serem utilizadas na rede neural deve ser feita de forma criteriosa, de modo a se obter um modelo mais rico em informações. Essa escolha é importante, uma vez que a presença de variáveis não relevantes pode afetar o treinamento da rede, comprometendo o seu poder de generalização.

Seixas et al. (1996) apresentaram uma proposta para identificação das variáveis explicativas mais relevantes, servindo como método para seleção das variáveis de entrada do modelo neural. O estudo da relevância de uma variável x_j estabelece uma comparação entre a saída obtida pela rede neural treinada (modelo final) e a resposta da rede neural obtida ao se manter fixa, esta mesma variável x_j, no seu valor médio, calculado sobre as amostras que compõem o conjunto de treino. Esse cálculo é feito para cada variável, segundo a equação:

$$R_j = \frac{1}{Npat}\sum_{i=1}^{Npat}\left[y(x_i, w) - \left(y(x_i, w) \mid x_{j,\,i} = \bar{x}_j\right)\right]^2 \qquad (4)$$

onde *Npat* é o número de padrões e $(y(x_i, w))$ corresponde à saída da rede. Quanto maior o valor obtido no cálculo da estatística Rj, maior será a relevância da variável.

A etapa seguinte do desenvolvimento do modelo consiste no treinamento da rede neural utilizando, como informação de entrada, apenas as variáveis indicadas como relevantes na etapa anterior e identificando, após o treinamento, o novo conjunto de variáveis que exerce maior influência na variável resposta. Com isto, pretende-se obter um modelo eficiente e mais compacto.

Neste trabalho, a análise de relevância foi feita adotando-se um limiar correspondente a, aproximadamente, 10% do maior valor encontrado na estatística Rj. Assim, foram consideradas relevantes as variáveis que tivessem apresentado Rj superior a este ponto de corte.

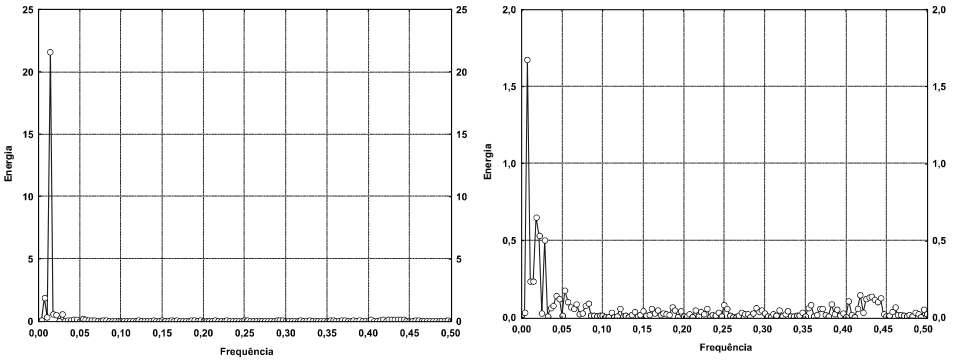
5 Resultados

A análise espectral da variável correspondente ao número de internações hospitalares (alvo de treino) foi realizada sobre a série normalizada. A Tabela 1 apresenta as frequências (normalizadas) mais relevantes da série alvo de treino, a energia correspondente e o valor resultante do cálculo da RMSE após a retirada do componente senoidal. Como se pode observar, a frequência de maior importância, na primeira análise espectral, foi 0,0143. Para avaliar a necessidade da retirada de algum componente senoidal remanescente, fez-se uma segunda análise espectral da série. As frequências mais relevantes foram 0,0071, 0,0179 e 0,0214. Os ciclos foram removidos mediante a observação do valor obtido no cálculo da RMSE. Notou-se também a presença de um componente senoidal com frequência 0,0286. Entretanto, a tentativa de remoção deste componente provocou o aumento do valor da RMSE (0,6889), motivo pelo qual este componente não foi retirado. Após a retirada dos componentes senoidais, observou-se a redução no valor da energia, inicialmente 21,6135, passando a 0,3539 ao final do processo.

Tabela 1 - Série das internações hospitalares – análise espectral

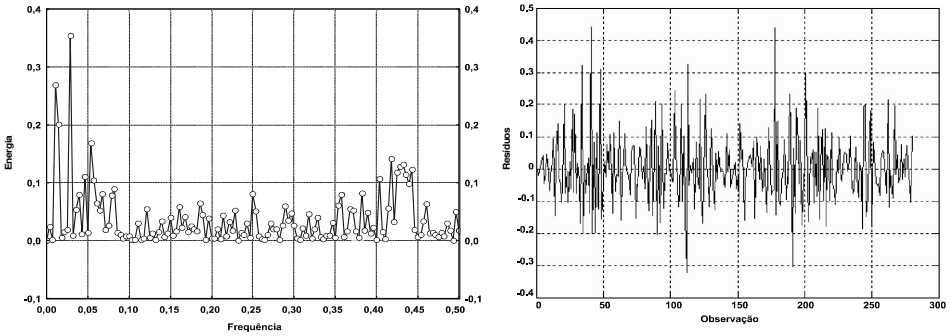
	Frequência	Energia	RMSE
1ª análise espectral	0,0143	21,6135	0,6949
2ª análise espectral	0,0071	1,6703	0,6896
	0,0179	0,6511	0,6882
	0,0214	0,3539	0,6870

A Figura 3 apresenta o espectro resultante da primeira análise e a Figura 4 mostra o espectro após a segunda análise. A Figura 4 mostra, ainda, a série de resíduos, que irá alimentar o modelo neural.



(a) (b)

Figura 3 - Série das internações hospitalares – análise espectral: espectro original (a) e resultante após a primeira análise (b).

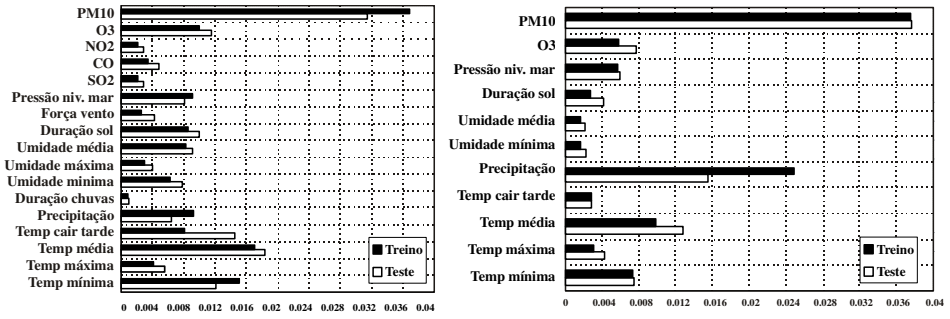


(a) (b)

Figura 4 - Série das internações hospitalares após a retirada dos componentes senoidais: espectro resultante (a) e série de resíduos (b).

Após as etapas de normalização e de pré-processamento, deu-se início à modelagem neural. A rede neural foi implementada sobre os resíduos obtidos após as etapas de normalização e pré-processamento dos dados, consistindo na retirada dos ciclos e, em seguida, das médias móveis de seis dias. O número de neurônios que compuseram a camada intermediária foi determinado de acordo com o menor valor calculado para o MAPE no conjunto de teste, após o pré-processamento dos dados. Foram desenvolvidos modelos neurais, com números distintos de neurônios na camada escondida, número esse que varia de 2 a 10 neurônios. O menor MAPE observado no conjunto de teste foi de 0,138. Assim, o modelo neural foi implementado com a topologia 17-7-1, representando 17 variáveis de entrada (12 relacionadas aos fatores climáticos e 5 referentes às concentrações de poluentes), 7 neurônios na camada intermediária e uma variável de saída (número de internações hospitalares).

Após a implementação do modelo neural, o critério Rj de relevância foi utilizado para reduzir a dimensionalidade da informação relativa aos fatores climáticos e de poluição atmosférica, indicando a informação mais relevante. A Figura 5(a) mostra o resultado da análise de relevância aplicada à saída do primeiro modelo neural implementado, onde de um total de 17 variáveis explicativas, 11 foram selecionadas.



(a) (b)

Figura 5 - Análise de relevância: (a) modelo completo e (b) variáveis selecionadas após a implementação do modelo completo.

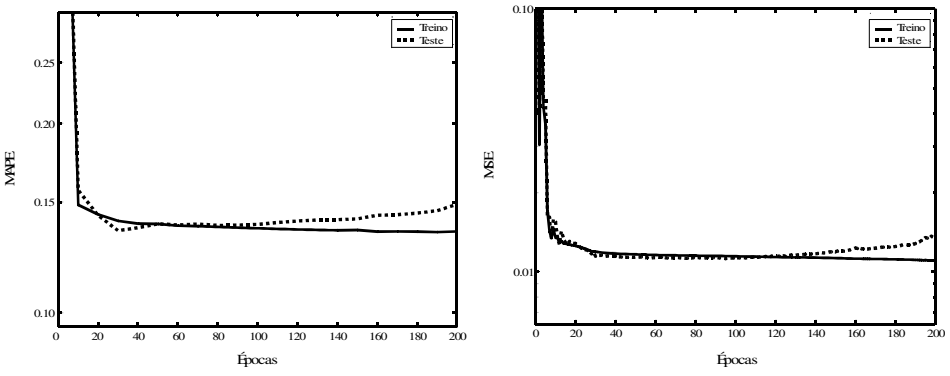
Uma nova redução de dimensionalidade foi obtida partindo-se desse modelo, cuja base de entrada era composta por essas 11 variáveis selecionadas, implementado com a topologia 11-5-1 (11 neurônios referentes às variáveis selecionadas pelo critério Rj; 5 neurônios na camada intermediária; e um neurônio na saída). Após uma nova aplicação do critério Rj sobre o modelo neural desenvolvido, 6 variáveis foram consideradas relevantes, sendo apresentadas na Figura 5(b): Material particulado com diâmetro inferior a 10μm (PM_{10}); Emissão de O_3; Pressão ao nível do mar; Precipitação diária; Temperatura média; e Temperatura mínima do dia. O modelo neural foi, então, projetado com a estrutura 6 – 4 – 1, representando 6 variáveis de entrada, 4 neurônios na camada intermediária e 1 na saída.

A Tabela 2 apresenta as medidas de desempenho dos modelos neurais implementados com suas respectivas topologias, onde se observa uma boa capacidade de generalização, com MAPE avaliado em torno de 0,13 no conjunto de teste, e uma boa qualidade no ajuste, com o MSE próximo de 1% no conjunto de treinamento.

Tabela 2 - Desempenho dos modelos neurais após o treinamento

Topologia	MSE (Treino)	MSE (Teste)	MAPE (Treino)	MAPE (Teste)
17 - 7 - 1	0,0087	0,0120	0,1355	0,1380
11 - 5 - 1	0,0105	0,0121	0,1342	0,1396
6 - 4 -1	0,0110	0,0112	0,1364	0,1377

A Figura 6 mostra o processo de aprendizagem durante o treinamento desta última rede neural. Observa-se a ocorrência do *overtraining* (Haykin, 2008) nas proximidades da época 100, que foi evitado através da parada prematura do treinamento na época 96.



(a) (b)

Figura 6 - Medidas de desempenho do modelo neural: MAPE (a) e MSE (b).

O estudo de relevância das variáveis explicativas deste modelo apontou o PM_{10} e a variável climática precipitação diária, como os principais responsáveis pelo número de internações nos hospitais de Paris, motivadas por bronquiolite infantil. A Figura 7 apresenta o resultado desta análise, mostrando a relevância das variáveis, tanto no conjunto de treinamento quanto no de teste.

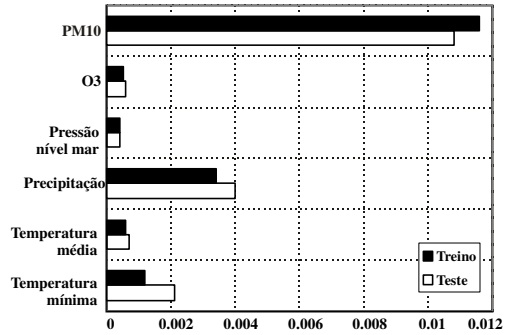


Figura 7 - Análise de relevância do modelo compacto.

Com a finalidade de desenvolver modelos neurais com a maior informação relevante possível, algumas redes neurais, com diferentes topologias, foram projetadas. Nesses novos modelos, foi adotado o mesmo tipo de pré-processamento que forneceu o menor valor calculado para o MAPE no conjunto de teste: retirada dos ciclos e, em seguida, a remoção das médias móveis de seis dias. As variáveis explicativas que compõem cada um desses modelos são as seguintes:

- Modelo 1: Somente variáveis climáticas;
- Modelo 2: Variáveis relevantes obtidas pelo critério Rj, para o Modelo 1;
- Modelo 3: Variáveis relevantes obtidas pelo critério Rj, para o Modelo 1, acrescentando-se os poluentes;
- Modelo 4: Variáveis relevantes obtidas pelo critério Rj, para o Modelo 1, acrescentando-se o PM_{10}, como única variável de poluição;
- Modelo 5: Variáveis climáticas selecionadas por Willems et al. (2007), através da análise de componentes principais: temperatura mínima do dia, precipitação diária, umidade média do dia, força do vento e pressão ao nível do mar;
- Modelo 6: Variáveis climáticas consideradas no Modelo 5, adicionando-se as variáveis relacionadas à poluição atmosférica;
- Modelo 7: Variáveis climáticas consideradas no Modelo 5, juntando-se o PM_{10}.

O estudo de Willems et al. (2007) apontou o PM_{10} como único responsável pelo número de internações hospitalares, conforme mencionado anteriormente, motivo pelo qual o poluente foi incluído como única variável de poluição nos modelos 4 e 7.

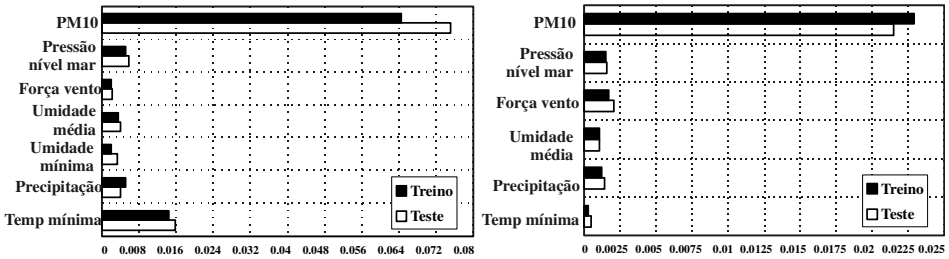
Após, cada implementação, o poder de generalização das redes neurais foi avaliado e a análise de relevância, segundo a estatística Rj, foi desenvolvida.

Nessas implementações, destacaram-se duas redes neurais (modelos 4 e 7), que diferem apenas pela inclusão da variável explicativa umidade mínima do dia na entrada de dados do modelo 4. A Tabela 3 apresenta as medidas de desempenho de ambos os modelos neurais, com suas respectivas topologias, onde se pode observar um bom poder de generalização e uma boa qualidade do ajuste, com valores próximos de 0,14 e 0,01, associados, respectivamente, ao MAPE (avaliado no conjunto de teste) e ao MSE (medido no conjunto de treinamento).

Tabela 3 - Desempenho das redes neurais para os modelos nos quais as bases de entrada diferem apenas pela presença da variável umidade mínima do dia

Modelo	Topologia	MSE (Treino)	MSE (Teste)	MAPE (Treino)	MAPE (Teste)
4	7 - 7 - 1	0,0103	0,0120	0,1434	0,1490
7	6 - 5 -1	0,0105	0,0115	0,1391	0,1478

A Figura 8 mostra a análise de relevância após o treinamento das referidas redes neurais, apontando o PM_{10} como variável mais relevante em ambos os modelos. Observa-se que quando a variável correspondente à umidade mínima do dia está presente, a rede neural identifica duas variáveis relevantes: PM_{10} e Temperatura mínima do dia. Observa-se, ainda, que na ausência da referida variável, o modelo neural aponta o PM_{10} como única variável relevante.



(a) (b)

Figura 8 - Análise de relevância: (a) variáveis climáticas obtidas pelo critério Rj, acrescentando-se o PM_{10} e (b) variáveis climáticas selecionadas pela análise de componentes principais, adicionando-se o PM_{10}.

Nos ensaios clínicos, há muitas fontes de variação que causam impacto no desfecho. As eventuais variações, que não puderem ser identificadas e controladas adequadamente, poderão interferir no efeito do tratamento em questão (Chow e Liu, 1995 apud Chow e Liu, 2004). No estudo da relação entre exposição e doença, um verdadeiro confundidor está associado à exposição em questão, sendo, ao mesmo tempo, um fator determinante da doença. Ou seja, ele confunde a relação entre a exposição e a doença (Vandenbroucke, 2002).

A inclusão da variável umidade mínima revelou, portanto, uma relação entre o número de internações hospitalares e a temperatura mínima do dia. Neste caso, a umidade mínima do dia pode ser considerada como uma variável de "anti-confundimento" (ou de confundimento negativo).

Conclusões

Os resultados obtidos através das redes neurais foram compatíveis com os encontrados por Willems et al. (2007), através da utilização dos modelos aditivos generalizados, apontando o material particulado com diâmetro inferior a 10μm (PM_{10})

como principal responsável no número de internações infantis nos hospitais de Paris devido a bronquiolite.

A metodologia utilizada neste estudo não é dependente das doenças do sistema respiratório, nem fica restrita a apenas uma determinada região geográfica, podendo ser estendida para outras aplicações.

Agradecimentos

Ao Professor Mounir Mesbah da Universidade Pierre et Marie Curie (Paris – França) por ter fornecido a base de dados que serviu como alicerce para o desenvolvimento deste estudo e à FAPERJ, ao CNPq e à CAPES pelo apoio, viabilizando a realização deste projeto.

NASCIMENTO, E. M.; PEREIRA, B. de B.; SEIXAS, J. M. Artificial neural networks: an application in the study of air pollution and its adverse health effects. *Rev. Bras. Biom.*, São Paulo, v.27, n.1, p.37-50, 2009.

- *ABSTRACT: There is a great need to assess the association between weather and air pollution with mortality or hospital admissions due to respiratory diseases. This paper proposes neural networks as alternative methodology to evaluate that association. The data refer to the number of hospitalizations in the city of Paris due to infant bronchiolitis, between 1997 and 2000. The neural models were evaluated for data description and to measure their capacity of generalization. The best results were obtained through data pre-processing, with removal of cycles and use of a moving average filter. A relevance study of the explanatory variables was also carried out. The results were consistent to those found through generalized additive models pointing out the particulate matter (PM_{10}) as the main responsible for the number of hospital admissions.*

- *KEYWORDS: Artificial neural network; air pollution; respiratory diseases.*

Referências

ALEKSEEV, K. P. G.; SEIXAS, J. M. A multivariate neural forecasting modeling for air transport – Preprocessed by decomposition: A Brazilian application, Journal of Air Transport Management, doi:10.1016/j.jairtraman.2008.08.008, p.1-5, 2008.

BATES, B. C.; KUNDZEWICZ, Z. W.; WU, S.; PALUTIKOF, J. P.(Ed.) Climate change and water. Technical paper of the intergovernmental panel on climate change, IPCC Secretariat, Geneva. 2008. 210p. Disponível em <http://www.ipcc.ch/pdf/technical-papers/climate-change-water-en.pdf> Acesso em 19 mar. 2009.

BRAGA, A.; PEREIRA, L. A. A.; SALDIVA, P. H. N. Poluição atmosférica e seus efeitos na saúde humana. In: SUSTENTABILIDADE NA GERAÇÃO E USO DE ENERGIA NO BRASIL: OS PRÓXIMOS VINTE ANOS, Campinas: UNICAMP, 2002.

CALÔBA, G. M.; CALÔBA, L. P.; SALIBY, E. Cooperação entre redes neurais artificiais e técnicas 'clássicas' para previsão de demanda de uma série de vendas de cerveja na Austrália. *Pesq. Oper.*, Rio de Janeiro, v.22, n.3, p.345-358, 2002.

CHOW, S. C.; LIU, J. P. *Design and analysis of clinical trials*: Concepts and Methodologies, 2.ed. New York: Wiley, 2004.

CHOW, S. C.; LIU, J. P. *Statistical design and analysis in pharmaceutical science*. New York: Dekker, 1995.

CONCEIÇÃO, G. M. S.; SALDIVA, P. H. N.; SINGER, J. M. Associação entre séries de mortalidade / morbidade e concentrações de poluentes atmosféricos: uma estratégia para análise estatística. *Rev. Bras. Estat*., Rio de Janeiro, v.63, n.219, p.75-98, 2002.

COOLEY, J. W.; TUKEY, J. W. An algorithm for the machine calculation of complex Fourier series. *Math. Comput*. v.19, p.297-301, 1965.

EVERARD, M. L.; SWARBRICK, A.; WRIGHTHAM, M. et al. Analysis of cells obtained by bronchial lavage of infants with respiratory syncytial virus infection. *Archives of Disease in Childhood*, v.71, p.428-432, 1994.

FARHAT, C. K.; CINTRA, O. A. L.; TREGNAGHI, M. W. Vacinas e o trato respiratório – o que devemos saber? *J Pediatr*, Rio de Janeiro, v.78 (Supl.2): S195-S204, 2002.

GUTIÉRREZ, A. M.; CASTELLANOS, E. V.; GARZA, A. Q. et al. Frecuencia de niños hospitalizados por el virus sincitial respiratorio en tres periodos invernales. *Rev Mex Pediatr*. v.70, n.4, p.167-170, 2003.

HASTIE, T. J.; TIBSHIRANI, R. J. *Generalized additive models*. London: Chapman & Hall, 1990.

HAYKIN, S. *Neural networks and learning machines*. Prentice Hall, 2008.

LIMA, L. P.; ANDRÉ, C. D. S.; SINGER, J. M. Modelos aditivos generalizados: metodologia e prática. *Rev. Bras. Estat*., Rio de Janeiro, v.62, n.217, p.37-69, 2001.

NELSON, M.; HILL, T.; REMUS, W.; O'CONNOR M. Times series forecasting using neural networks: should the data be deseasonalized first? *J. Forecast*., v.18, p.359-367, 1999.

PEREIRA, B. B.; RODRIGUES, C. V. S. *Redes neurais em estatística*, 13º SINAPE, 144p., Caxambu, 1998.

RIEDMILLER, M.; BRAUN, H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, 1993, San Francisco. *Proceedings*… p.586-591.

RIEDMILLER, M. *RPROP – description and implementation details*. Technical Report. Universitat Karlsruhe, 1994. Disponível em:
<http://citeseer.ist.psu.edu/riedmiller94rprop.html>. Acesso em 20 mar. 2009.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R.J. *Learning internal representations by error propagation*. In: PARALLEL DISTRIBUTED PROCESSING, MIT Press, Cambridge, MA, v.1., 1986.

SEIXAS, J. M.; CALÔBA, L. P.; DELPINO, I. *Relevance criteria for variance selection in classifier designs*. In: INTERNATIONAL CONFERENCE ON ENGINEERING APPLICATIONS OF NEURAL NETWORKS, 1996. *Proceedings*...p.451-454.

SINGER, J. M.; ANDRÉ, C. D. S.; LIMA, L. P.; CONCEIÇÃO, G. M. S. *Association between atmospheric pollution and mortality in São Paulo, Brazil*: regression models and

analysis strategy. In: INTERNATIONAL CONFERENCE ON STATISTICAL DATA ANALYSIS BASED ON THE L1 NORM AND RELATED METHODS, 4, 2002, Neuchâtel. *Proceedings...* p.429-450.

ŠRÁM R. J.; BINKOVÁ, B.; DEJMEK, J. et al. *Intrauterine growth retardation, low birth weight, prematurity and infant mortality*. In: EFFECTS OF AIR POLLUTION ON CHILDREN'S HEALTH AND DEVELOPMENT: A REVIEW OF THE EVIDENCE. Bonn: WHO European Centre for Environment and Health. p.14-27, 2005.

UNFCCC - United nations framework convention on climate change, 1992. Disponível em: <http://unfccc.int/resource/docs/convkp/conveng.pdf>. Acesso em: 12 dez. 2008.

UNFCCC - United nations framework convention on climate change. *Kyoto protocol reference manual on accounting of emissions and assigned amounts*, 2007. Disponível em: <http://unfccc.int/files/national_reports/accounting_reporting_and_review_under_the_kyoto_protocol/application/pdf/rm_final.pdf>. Acesso em: 12 dez. 2008.

VANDENBROUCKE, J. P.The history of confounding. Soz. - Und Präventivmed. *Soc. Prev. Med.*, v.47, n.4, p.216-224, 2002.

WILLEMS, S.; SEGALA C.; MAIDENBERG, M.; MESBAH, M. *Longitudinal analysis of short-term bronchiolitis air pollution association using semiparametric models*. In: ADVANCES IN STATISTICAL METHODS FOR THE HEALTH SCIENCES, 2007. Auget, J. L.; Balakrishnan, N.; Mesbah, M.; Molenberghs, G. (Eds.). Boston: Birkhäuser, Springer, 2007. p.467-487.

ZANOBETTI, A.; WAND, M. P.; SCHWARTZ, J.; RYAN, L. M. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, New York, v.1, n.3, p.279-292, 2000.

Recebido em 15.12.2008.

Aprovado após revisão 06.04.2009.

Apêndice F

Modelos Log-lineares

Os modelos log-lineares foram utilizados no artigo intitulado “Prevalência de alterações contráteis segmentares e sua relação com arritmias ventriculares complexas em pacientes chagásicos com eletrocardiograma normal ou bordeline”.

O artigo foi aceito, em julho de 2010, para publicação na “Revista da Sociedade Brasileira de Medicina Tropical”.

Título: Prevalência de alterações contráteis segmentares e sua relação com arritmias ventriculares complexas em pacientes chagásicos com eletrocardiograma normal ou “bordeline”.

Title: Regional left ventricular dysfunction prevalence and the relation with complex ventricular arrhythmia in chagasic patients with normal or “borderline” electrocardiogram.

Autores: Flavia Vernin de Oliveira Terzi[1], Roberto Coury Pedrosa[1], Aristarco Gonçalves de Siqueira Filho[1], Emília Matos do Nascimento[4], Basílio de Bragança Pereira[5].

4- Programa de Engenharia de Produção – COPPE/UFRJ – E-mail: emilia@pep.ufrj.br.

5- Professor Titular de Bioestatistia e Estatistica Aplicada, Faculdade de Medicina e COPPE-Programa de Pos-Graduação da Escola de Engenharia-UFRJ e Hospital Universitario Clementino Fraga Filho-UFRJ.

RESUMO

Introdução: Morte súbita representa principal causa de óbito em chagásicos. Eventos fatais em pacientes saudáveis, com anormalidades contráteis, foram documentados.

Objetivos: Determinar a relação entre alteração contrátil e arritmias ventriculares complexas em pacientes chagásicos assintomáticos, classe funcional I e eletrocardiograma normal ou bordeline.

Métodos: 49 pacientes com Doença de Chagas e eletrocardiograma normal ou borderline realizaram ecocardiograma, teste ergométrico e Holter. Avaliou-se a contratilidade global e segmentar dos ventrículos e a presença de arritmias ventriculares complexas induzidas no esforço e espontâneas, respectivamente. A análise estatística foi feita pelo modelo Log-Linear geral.

Resultados: Idade média de 56 anos; 55% mulheres. Alterações contráteis segmentares em 24,5% dos pacientes; 12% dos Holter e 18% dos testes ergométricos positivos. Houve relação entre arritmia e alteração segmentar condicionada à presença da disfunção sistólica leve do ventrículo esquerdo.

Conclusões: Alterações contráteis, na presença de disfunção global leve, indicam pacientes sob maior risco de arritmias complexas.

Palavras-chave: Doença de Chagas, eletrocardiograma normal ou bordeline, alteração segmentar, arritmias ventriculares complexas.

ABSTRACT

Background: Sudden death is the major cause of death among chagasic. Significant number of fatal events in patients without apparent heart disease and “borderline” electrocardiogram, but with contractile ventricular dysfunction, has been documented.

Objectives: Determine the relation between regional dysfunction and the presence of ventricular arrhythmia in chagasic patients without apparent heart disease.

Methods: 49 patients with normal or “borderline” electrocardiogram were submitted to echocardiogram, exercise stress test and Holter. It was analyzed the presence of cardiac contractile alterations and complex ventricular arrhythmia. Statistic analysis used general “Log-Linear” model.

Results: Mean age 56 years; 55% women. Regional ventricular dysfunction in 24,5% of patients; Positive Holter in 12% and exercise stress test in 18%. There were relation between complex ventricular arrhythmia and contractile abnormalities in the presence of mild left ventricle dysfunction.

Conclusion: Regional contractile abnormalities with mild left ventricle dysfunction in Chagasic patients indicate a group with higher risk of complex ventricular arrhythmias, who needs for specific follow-up.

Key-words: Chagas’ Disease, normal or borderline electrocardiogram, regional contractile abnormalities, complex arrhythmias.

INTRODUÇÃO

A Doença de Chagas é considerada a quarta moléstia de maior impacto médico-social na América Latina.[1] Tem sido amplamente combatida no Continente Americano nas ultimas décadas. No Brasil, ainda constitui um dos maiores problemas de saúde pública.[2]

A doença tem fases aguda e crônica, caracterizada por sinais e sintomas distintos separados por um período indeterminado, porém representados por uma miocardite discreta a moderada que pode persistir por anos sem manifestações clínicas.[2] A forma crônica da doença ocorre em aproximadamente 70% dos casos, pode durar de 10 a 30 anos mas, na maioria das vezes, persiste por toda vida, com uma fisiopatologia de curso variável e intrigante.[3]

Os pacientes com a forma indeterminada pertencem a categoria de potenciais cardiopatas, uma vez que a cada ano aproximadamente 3% destes terão comprometimento cardíaco.[4]

Pacientes com cardiopatia chagásica que sofrem de insuficiência cardíaca tem um alto risco para desenvolver morte súbita elétrica. No entanto, a freqüência de morte súbita nos pacientes com função ventricular preservada também é expressiva. Portanto, a prevenção deve ser feita de forma abrangente.[5]/[6]/[7]

Alguns estudos sugerem um aumento na mortalidade em pacientes com alterações “borderline” no eletrocardiograma.[8] Porém, não há um consenso para a melhor estratégia para estratificação de risco para prevenção primária de morte súbita elétrica em pacientes chagásicos sem cardiopatia aparente (Classe Funcional I da NYHA).

A incidência de morte súbita elétrica nos pacientes com disfunção

ventricular é maior do que naqueles que têm função preservada, mas considerando-se em números absolutos, a prevalência de morte súbita elétrica é maior entre aqueles pacientes com função ventricular preservada, uma vez que estes são em número significativamente superior.

A real incidência de morte súbita elétrica em pacientes com Doença de Chagas sem disfunção ventricular não está totalmente esclarecida. O manejo desses pacientes objetiva a predição da morte súbita elétrica por meio da identificação de possíveis marcadores fisiopatológicos.[9,10]

O presente estudo investiga a associação entre anormalidade contrátil segmentar e o perfil arritmogênico ventricular dos pacientes chagásicos crônicos com o eletrocardiograma normal ou “bordeline” através da avaliação cardíaca por meio de exames não-invasivos: o ecocardiograma de repouso, o teste de esforço e o Holter.

Material e Método

Foram incluídos pacientes com idade entre 20 a 70 anos admitidos no ambulatório de cardiopatia chagásica do Serviço de Cardiologia do Hospital Universitário da UFRJ com diagnóstico de Doença de Chagas, afastados da zona endêmica há mais de 20 anos que tiveram o eletrocardiograma de repouso normal ou borderline. Pacientes com os seguintes critérios foram excluídos: não completaram o protocolo inicial de admissão; outras doenças cardíacas associadas; pacientes com cardiopatia isquêmica confirmada após investigação; pacientes tratados especificamente para Doença de Chagas, obesidade mórbida, hipertenso com lesão de órgãos-alvo, diabetes melito de longa data, doença renal

crônica nos estágios III a V; doença pulmonar obstrutiva crônica (todas as formas); tabagismo, alcoolismo e doenças sistêmicas importantes.

O projeto teve aprovação do Comitê de Ética em Pesquisa local, sob Parecer CEP n° 208/07, atendendo as Diretrizes Éticas Nacionais e Internacionais para Pesquisa Biomédicas Envolvendo Seres Humanos.[11]

Protocolo de Avaliação

Exame Clínico

Após confirmação sorológica, todos os pacientes foram submetidos a uma avaliação clínica inicial padronizada (anamnese, exame físico, exame laboratorial, eletrocardiograma de repouso e radiografia de tórax). A avaliação laboratorial incluiu: hemograma completo, glicemia, uréia e creatinina, ácido úrico, lipidograma, sódio, cálcio e potássio, proteínas totais e frações, provas de função hepática e tireoideana, exame parasitológico de fezes e análise de urina para elementos anormais e sedimentos.

Todos os pacientes pertencentes ao presente estudo realizaram, no período de Janeiro a Dezembro de 2007, os seguintes exames: eletrocardiograma convencional de repouso com 12 derivações, monitorização eletrocardiográfica ambulatorial dinâmica de 24 horas (sistema Holter), ecocardiograma uni e bidimensional com Doppler e teste de esforço. O intervalo entre os exames foi de no máximo 30 dias. O uso de fármacos dromo ou cronotrópicos negativos ou hipotensores (como digital, betabloqueador, vasodilatador ou antiarrítmicos) era interrompido 48 horas antes das avaliações clinicas e laboratoriais.

Eletrocardiograma

O eletrocardiograma (ECG) foi analisado conforme os critérios da NYHA, utilizando-se o código de Minnessota[12] modificado para a cardiopatia chagásica.[13]

Foram considerados: *Normal* - Ausência de qualquer alteração listada sob a designação “borderline” com ritmo sinusal normal. Os seguintes critérios foram aceitáveis: onda R apiculada em V5 se não associada com alteração da onda T ou segmento ST e Rr’ in V1 na ausência de outros critérios para bloqueio incompleto de ramos direito. *Borderline* - Ausência de qualquer alteração listada sob a designação “normal” mais um ou mais do seguinte: alteração mínima da onda Q; desvio para direita do complexo QRS na ausência de outros critérios para bloqueio completo do ramo direito; Ondas R precordiais apiculadas sem alteração da onda T ou do segmento ST; alterações inespecificas do segmento ST e/ou da Onda T; arritmia sinusal ou taquicardia sinusal; bradicardia sinusal; extra-sistole atrial ou juncional não repetitiva; ritmo juncional e onda p alargada se associada com taquicardia sinusal.

Eletrocardiograma dinâmico contínuo (sistema Holter)

A análise do traçado eletrocardiográfico do Holter foi realizada de maneira prospectiva. O analisador foi solicitado a classificar os batimentos como normais ou ectópicos à medida que o sistema processava a informação. A leitura do Holter foi feita duas vezes, em momentos diferentes, pelo mesmo observador.

A instabilidade elétrica ao Holter de 24 horas foi definida pela presença de arritmias ventriculares complexas (extra-sístoles ventriculares > 30/hora), episódios de taquicardia ventricular monomórfica sustentada (definida como mais de 3 batimentos consecutivos com duração > 30 segundos ou mais) ou episódios de taquicardia ventricular monomórfica não sustentada (definida como 3 ou mais batimentos consecutivos com duração < 30 segundos);

Ecocardiograma uni e bidimensional com Doppler:

Todos os exames foram feitos pelo mesmo ecocardiografista. Foram escolhidos aleatoriamente alguns exames para serem gravados e avaliados posteriormente por outro ecocardiografista.

A avaliação foi feita de acordo com os critérios preconizados pela *American Society of Echocardiography,*[14] A função sistólica global do VE foi quantificada através do cálculo da fração de ejeção do ventrículo esquerdo pelo método de Teicholz e Kreulen, sendo classificada em: normal (fração de ejeção >= 55%), levemente deprimida (fração de ejeção < 55% e >=45%), moderadamente deprimida (fração de ejeção <45% e >=35%) ou gravemente deprimida (fração de ejeção < 35%). A função diastólica foi avaliada através da análise do enchimento ventricular esquerdo pelo fluxo mitral[15/16] e Doppler tecidual do anel mitral lateral[17/18].

Ecocardiograma anormal foi definido pela presença de qualquer anormalidade contrátil segmentar (hipocinesia, acinesia ou discinesia) e/ou disfunção ventricular sistólica.

Teste de Esforço

O teste de esforço foi realizado em esteira de marca Funbec. Utilizou-se o protocolo de Bruce modificado. Foram considerados para análise apenas os testes de esforço dos pacientes que atingiram 7 MET, limite definido como nível submáximo de exercício (boa correlação entre o limiar anaeróbio atingido e a carga atingida de 7 MET).[19] Foi valorizada a análise das arritmias ventriculares presentes com suas respectivas morfologias e taquicardia ventricular.

Análise estatística

Este estudo foi realizado sob a consultoria estatística da Comissão de Investigação Científica (CIC) da Divisão de Pesquisa (DPq) do HUCFF/UFRJ.

As variáveis obtidas formam armazenadas num banco de dados do programa Microsoft Office Excell 2003, de onde foram coletados dados para a criação de um banco de dados do Sistema R. A análise dos dados, para detectar o padrão de dependência entre as variáveis estudadas e o quanto de cada associação era devido á influência dos outros fatores, eliminando o confundimento de efeitos, foi realizada pelo modelo log-linear geral [20] e na análise univariada foi utilizado o Teste de Fisher.

Os dados demográficos e as medidas ecocardiográficas foram analisados através da média, mediana e desvio padrão.

A concordância intra e inter-observador para o diagnóstico de anormalidade contrátil segmentar e disfunção ventricular ao ecocardiograma assim como das arritmias ventriculares induzidas no esforço foram mensuradas pelo método estatístico Kappa. Foi utilizada a classificação proposta por Landis e Koch.[21]

RESULTADOS

Dados clínicos e Demográficos

Do grupo de 308 pacientes chagásicos que são acompanhados no Ambulatório de Cardiologia do HUCFF, 62 preencheram os critérios definidos para inclusão no trabalho. Doze pacientes não realizaram um ou mais exames, tendo sido excluídos da análise estatística e um paciente foi vítima de morte súbita antes de finalizar o projeto (só tinha realizado Holter de 24h, com laudo normal). Seguiram todo o protocolo de exames no período estabelecido o total de 49 pacientes.

A média de idade do grupo estudado foi de 56,4 ± 12 anos (mediana 58 anos) sendo 57% da amostra constituída por mulheres. As características clínicas dos pacientes encontram-se na Tabela 1.

Tabela 1: Características Gerais da População

Número de pacientes estudados	49
Idade (anos)	56,4±12,48
Sexo masculino	21
Sexo feminino	28
Tempo para o diagnóstico (meses)	218,98±152,62
Manifestações clínicas da doença	11
Megaesôfago	6
Megacólon	3
Megaesôfago e megacólon	2
Doenças associadas	
Hipertensão arterial	4
Dislipidemia	2
Diabetes	2
Esôfago de Barret	2
Dor precordial	1
Doença cérebro-vascular	1
Vasculopatia periférica	1
Outras	10
Drogas em uso	
IECA	25
Betabloqueador	17
Diurético	4
Antagonista do cálcio	2
Outros	4

IECA: enzima conversora da angiotensina.

1. Avaliação Ecocardiográfica

Foram identificadas alterações contráteis segmentares em 24,5% dos pacientes (n=12). Destes, 25% (n=3) tinham comprometimento leve da função global do ventrículo esquerdo, 25% (n=3) tinham disfunção global leve do ventrículo direito e disfunção leve biventricular em 8%.

Em relação às paredes acometidas pelo déficit segmentar, dois pacientes apresentaram alterações em múltiplos seguimentos. Um paciente apresentava

acinesia ínfero-posterior, lateral e comprometimento do ventrículo direito. O outro apresentava acinesia das paredes inferior e ântero-septal. Dos dez pacientes restantes, 60% (n= 6) apresentaram acinesia ínfero-posterior, 30% acometimento do ventrículo direito e 10% discinesia apical. Foi identificado aneurisma em dois pacientes (17%), um acometendo ponta do ventrículo esquerdo e outro o segmento ínfero-basal.

2. Avaliação do Holter

Dos exames de Holter avaliados, 12% (n=6) apresentavam alterações. Um paciente só apresentou extra-sístoles ventriculares (30/h). O restante apresentou tanto taquicardia ventricular (variando de 1 a 30 episódios não-sustentadas) quanto extra-sístoles ventriculares (variando de 60 a 289/h).

3. Avaliação do teste ergométrico

No teste ergométrico ocorreram alterações em 18% dos pacientes (n=9). Seis pacientes apresentavam extra-sístoles ventriculares no repouso e evoluíram com aumento no número de arritmias ventriculares (extra-sístoles ventriculares e taquicardia ventricular monomórfica não sustentada) durante o esforço. Três pacientes tinham eletrocardiograma de repouso sem arritmia ventriculares. Destes, dois apresentaram extra-sistolia ventricular e um apresentou episódios de taquicardia ventricular monomórfica não sustentada durante o esforço.

Fatores como idade, gênero, uso de medicações cardiológicas e presença de disfunção diastólica não mostraram significância estatística em relação à ocorrência de arritmias, tanto espontâneas ou induzidas (Tabela 2).

Tabela 2 – Teste de Fisher

	Valor de p (λ^2)
Tempo de Diagnóstico x AS	0,512

Tempo de Diagnóstico x Holter (Arritmia)	0,133
Tempo de Diagnóstico x TE (Arritmia)	0,185
Gênero x FSG	1
Gênero x AS	1
Gênero x Holter (Arritmia)	1
Gênero x TE (Arritmia)	0,714
Idade x Função Global	1
Idade x AS	0,182
Idade x Holter (Arritmia)	0,388
Idade x TE (Arritmia)	0,488
Arritmia x Uso de beta-bloqueador	0,467
Arritmia x Uso de IECA	0,098
Alteração segmentar x TE (c/ e s/disfunção)	**0,004**
Alteração segmentar x TE (s/disfunção)	0,211
Alteração segmentar x Holter (c/ e s/disfunção)	**0,0001**
Alteração segmentar x Holter (s/disfunção)	**0,028**
Arritmia x Disfunção Diastólica	1

4. Confiabilidade da classificação de arritmias ventriculares esforço induzidas e alterações contráteis segmentares.

Na concordância inter-observador para detecção de arritmias ventriculares no repouso e no esforço (15 minutos cada período) obteve-se um valor Kappa de 0,85 (IC 95%=0,72-0,92). A concordância intra-observador teve um valor kappa de 0,95 (IC 95%=0,74-0,99).

A concordância inter-observador para detecção de alterações contráteis segmentares nos registros ecocardiográficos em repouso foi realizada em 16 pacientes, obtendo-se um valor Kappa de 0,35 (IC 95%=0,207-0,901). A concordância intra-observador teve um valor kappa de 0,71 (IC 95%=0,332-1,086).

5. Modelo Log-linear

Devido à existência de múltiplas variáveis com potencial relação com a presença de arritmias malignas e alterações segmentares ao ecocardiograma, com graus de dependência variados entre elas, utilizou-se a análise multivariada seguindo o modelo Log-Linear geral (Figura 16 e Tabela 6). Identificaram-se as seguintes relações entre as variáveis: a linha pontilhada indica uma associação mútua entre idade, arritmia e disfunção global do VE ao nível de 0.10. Observe que arritmia e alteração segmentar são independentes condicionadas à disfunção global do VE, ou seja, indica dependência apenas na presença de disfunção leve.

Tabela 6: Resultados do modelo log linear para a avaliação da relação entre as variáveis estudadas

NULL	P (> | Chi |)
Idade	0,014
FSG	2,269e-10
AS	2,549e-04
DD	0,886
Arrit	4,158e-06
Idade: FSG	0,453
Idade: AS	0,508
FSG: AS	2,041e-04
Idade: DD	3,100e-06
FSG: DD	0,529
AS : DD	0,941
Idade: Arrit	0,100
FSG: Arrit	0,066
AS: Arrit	0,145
DD: Arrit	0,343
Idade: FSG: Arrit	0,073

FSG - Disfunção sistólica leve de VE, AS – Alteração segmentar, DD – Disfunção diastólica, Arrit arritmia ventricular complexa.

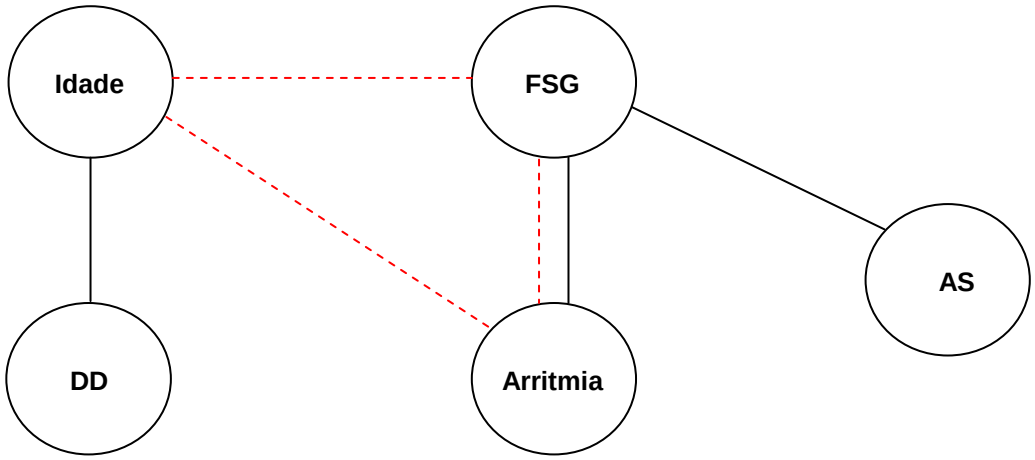


Figura 16: Gráfico após análise estatística pelo modelo Log-Linear para p=0,10.

Idade (0 se < 54; 1 caso contrário), FSG - Disfunção sistólica global leve do VE (0 = Normal / 1 = Disfunção leve), AS - Alteração segmentar (0 = Ausente / 1 = Presente), DD - Disfunção diastólica (0 = Ausente / 1 = Presente), Arritmia (0 = Negativo / 1 = Positivo).

DISCUSSÃO

A doença de Chagas permanece como relevante problema de saúde pública e causa fundamental de cardiomiopatia secundária e de morte súbita no país.

Muita controvérsia existe sobre quais fatores poderiam influenciar a evolução da doença de Chagas, já que alguns pacientes evoluem para formas mais graves, enquanto outros permanecem assintomáticos por toda a vida. Um maior entendimento dos fatores prognósticos da doença de Chagas tem ocorrido ao longo dos anos. Já é possível reconhecer fatores preditores independentes e, consequentemente, pacientes sob diferentes níveis de risco, facilitando e direcionando o tratamento. A arritmia cardíaca, freqüentemente ligada à disfunção ventricular, tem valor prognóstico, tendo pior evolução àqueles com arritmias ventriculares complexas e maior grau de disfunção ventricular. No entanto, ainda é desconhecido o valor das arritmias cardíacas na presença de alterações segmentares contráteis com disfunção ventricular leve.

Anormalidades contráteis do ventrículo esquerdo, mesmo que mínimas, parecem ter implicações prognósticas. Resultados preliminares [22] mostram que pacientes chagásicos com alterações segmentares contráteis do ventrículo esquerdo no ecocardiograma basal, seguidos por um período médio de 4,6 anos, apresentam progressão mais rápida da disfunção ventricular do que aqueles sem alterações de mobilidade das paredes ventriculares. Segundo os autores, no contexto de uma doença inflamatória como a doença de Chagas, deve-se considerar o portador de alterações segmentares como sendo um cardiopata, inferindo que tais pacientes são potenciais candidatos para morte súbita.

A análise da função ventricular segmentar, através da avaliação da motilidade e/ou espessamento dos segmentos das paredes ao ecocardiograma, tem caráter subjetivo e está relacionada a variações inter e intra-observador, uma vez que a experiência requerida na quantificação subjetiva da motilidade miocárdica demanda tempo de experiência com o método. A utilização de técnicas que permitam a análise quantitativa do estado contrátil do miocárdio oferece grande potencial no estudo da função cardíaca. O Doppler tissular apresenta-se como opção para a avaliação da contratilidade regional. Em estudo com 77 pacientes (controle normal, chagásicos com eletrocardiograma normal e chagásicos com eletrocardiograma alterado), as alterações pelo Doppler tissular estiveram presentes em ambos os grupos de pacientes chagásicos,[23] comparativamente ao grupo controle, denotando um comprometimento precoce da função sistólica nesses pacientes, mesmo na ausência de anormalidades eletrocardiográficas. Portanto, o uso de recursos ecocardiográficos mais avançados e certamente um número maior de pacientes poderia tornar os valores de concordância inter-observador encontrados nesta

dissertação aceitáveis, diminuindo o caráter subjetivo e aumentando a confiabilidade no método.

Em grande casuística[Erro! Indicador não definido.] com perfil de pacientes semelhante aos do presente estudo, 13% apresentaram alteração segmentar e 0,8% tinham comprometimento da função sistólica global do ventrículo esquerdo. Na presente casuística identificou-se 24,5% de pacientes com alterações segmentares, sendo que 6% (n=3) do total de pacientes apresentavam disfunção global do ventrículo esquerdo, 6% (n=3) do total de pacientes evoluíram com disfunção do ventrículo direito e 2% (1 paciente) apresentavam disfunção biventricular. Nossa prevalência foi maior do que as referências, provavelmente devido a diferenças no que se referem o delineamento do estudo e nas definições usadas.

Acredita-se que o substrato arrítmico primário (arritmias ventriculares espontâneas) define o risco de óbito em pacientes chagásicos cardiopatas, enquanto as arritmias ventriculares induzidas pelo esforço classificam os pacientes sob risco de morte súbita. No presente trabalho foi observado que a arritmia cardíaca complexa correlaciona-se com alteração de contratilidade segmentar condicionada à presença de disfunção ventricular leve de VE. Por outro lado, sabe-se que a incidência de morte súbita neste grupo de pacientes em valor absoluto é significativa (0,04 para cada 1000 pacientes-ano numa população de 3 milhões).[24] Podemos inferir, de modo indireto, que a detecção de arritmias cardíacas complexas neste grupo de pacientes possa ter alguma importância na estratificação para morte súbita.

O dado mais expressivo deste estudo é que pacientes chagásicos crônicos com eletrocardiograma normal ou bordeline tiveram uma prevalência significativa de alterações segmentares contráteis e estas somente se relacionaram com arritmias

ventriculares complexas condicionadas à presença de disfunção ventricular sistólica leve. A análise dos dados pelo modelo Log-linear mostrou uma forte relação entre a disfunção ventricular leve e as alterações segmentares. Também foi observada uma interdependência entre a disfunção ventricular leve e a presença de arritmias ventriculares complexas e esta relação era totalmente independente da presença de disfunção diastólica ou da idade.

Sendo assim, pode-se imaginar um cenário em que as várias hipóteses formuladas para explicar a instabilidade elétrica ventricular em pacientes chagásicos com eletrocardiograma normal ou bordeline trabalham de forma sinérgica para a produção de um mesmo quadro clínico, a disfunção ventricular sistólica.

Implicação clinica

Na doença de Chagas, todos os pacientes com evidência de comprometimento estrutural cardíaco (seja no eletrocardiograma, na radiografia de tórax ou no ecocardiograma) devem ser considerados potencialmente sob risco de morte súbita. Recentemente, em pacientes chagásicos ambulatoriais, Rassi e cols estimaram a incidência de morte cardíaca súbita em torno de 24 por 1000 pacientes-ano, taxa esta substancialmente maior em relação à população geral. Assim, a prevenção da morte arrítmica na doença de Chagas na sua forma cardíaca representa, hoje em dia, um grande desafio.

Por outro lado, sabe-se que a morte súbita é uma complicação rara de uma condição comum, a forma indeterminada (chagásicos sem cardiopatia aparente) da doença de Chagas. A prevenção de morte súbita é fundamental para a redução da mortalidade geral e principalmente neste grupo populacional, uma vez que a causa-mortis desses pacientes é devido principalmente à morte arrítmica.**Erro! Indicador não**

definido./Erro! Indicador não definido. Trabalhos apontam que a morte súbita neste grupo de pacientes acontece, geralmente, durante as atividades rotineiras, de esforço físico ou estresse emocional, nos chagásicos aparentemente saudáveis sem nenhum conhecimento ou indício sintomático de doença cardíaca.Erro! Indicador não definido./Erro! Indicador não definido./Erro! Indicador não definido. Outros apontam que este grupo de pacientes com morte súbita ressuscitada são mal respondedores à terapia atual, mesmo após implante de cardioversor-desfibrilador intracavitário.Erro! Indicador não definido./25

Assim sendo, o principal achado desse estudo pode contribuir de maneira importante para a prática clínica no manuseio dos pacientes chagásicos. A identificação precoce de alterações segmentares contráteis associadas à disfunção ventricular sistólica leve em pacientes chagásicos sem alterações eletrocardiográficas significativas, pode direcionar esse grupo de pacientes para análise mais criteriosa, como teste ergométrico e Holter de 24h, com o objetivo de estratificar o risco de arritmias complexas e potencial complicação em uma fase precoce da doença de Chagas. É necessário avaliar o custo-benefício e a acessibilidade de cada método (teste ergométrico e Holter de 24h) na estratificação do risco.

A necessidade de continuar a busca por um perfil de risco individual não limitado a extensão do comprometimento miocárdico se impõe neste grupo de pacientes, como a combinação com outros marcadores de risco tais como aqueles que refletem a disfunção autonômica, o perfil do estresse oxidativo e a presença de inflamação subclínica.

Em adultos na forma crônica, embora a negativação sorológica ocorra apenas numa minoria dos pacientes,[26] o tratamento etiológico parece ser eficaz na prevenção da progressão da cardiopatia chagásica. Em um ensaio terapêutico

controlado com chagásicos[27] seguidos, em média, oito anos, o tratamento específico com benzonidazol diminuiu o aparecimento de novas lesões eletrocardiográficas (4,2% x 30%) no grupo tratado em relação ao controle, diminuindo também a freqüência de pacientes com deterioração clínica cardíaca (2,1% x 17%). Portanto, existem evidências de que o tratamento etiológico pode prevenir a progressão da doença de Chagas, indicando que é hora de redefinir o papel do tratamento específico na forma crônica. É também plausível admitir que nos pacientes com mínimas alterações da função ventricular regional seja possível testar o real valor do tratamento etiológico da doença de Chagas. São os que potencialmente mais poderão se beneficiar com tal tipo de tratamento, se a teoria de que a ação inflamatória (diretamente relacionada à persistência parasitária, e mediada e/ou exaltada por agressão auto-imune) mostrar-se correta pela evidência indireta de cunho terapêutico obtido.

A conduta frente ao chagásico na forma crônica não pode continuar sendo a observação passiva dos pacientes. Estratégias no sentido de se definir grupos de risco, sujeitos a intervenção terapêutica e/ou readaptação profissional, devem ser estabelecidas, utilizando-se dados clínicos, epidemiológicos e aqueles obtidos através da avaliação cardíaca não-invasiva.

Limitações do estudo

No presente trabalho, apesar de não podermos afastar definitivamente o diagnóstico de insuficiência coronariana como um fator de confundimento importante, dados clínicos (fatores de risco representado por hipertensão e diabetes mellitus em número pequeno de pacientes, ausência sinais e sintomas de doenças cardiovasculares), dados eletrocardiográficos e teste de esforço nos permitem inferir pela ausência de doença coronariana obstrutiva funcionalmente significativa.

Baseados em trabalhos publicados sobre o estudo cineangiocoronariográfico nesse grupo de pacientes, com invariavelmente coronárias normais, achou-se pertinente não realizar cinecoronariografia.

Neste trabalho não foi testada a reprodutibilidade da arritmia esforço induzida uma vez que esta variável foi motivo de observação. Dados na literatura são inexistentes. Entende-se a necessidade da realização do mesmo.

CONCLUSÃO

- Não existe associação entre anormalidades contráteis segmentares e arritmias cardíacas complexas espontâneas e esforço induzida. A relação de interdependência está condicionada à presença de disfunção ventricular sistólica leve.

Bibliografia

[1] Salvatella R. Achievements in controlling Chagas disease in Latin América. Conference in Geneva (WHO), July 6, 2007.

[2] Marin-Neto JA, Simões MV, Sarabanda AV. Chagas' heart disease. Arq Bras Cardiol 1999; 72(3):247-80

[3] Macedo V. Indeterminate form of Chagas' disease. Mem Inst Oswaldo Cruz 1999; 94 (Suppl I):311-16

[4] Dias JC. The treatment of Chagas disease (South American Trypanosomiasis). Annals of Internal Medicine 2006; 144:772-74.

[5] Rassi A Jr, Rassi A, Rassi SG. Predictors of mortality in chronic Chagas disease: A systematic review of observational studies. Circulation 2007; 115:1101-08.

[6] Sternick EB, Martinelli M, Correia S R, Gerken LM, Teixeira RA, Scarpelli R, Scanavacca M, Nishioka SD, Sosa E. Sudden cardiac death in patients with Chagas heart disease and preserved left ventricular function. J Cardiovasc Electrophysiol 2006; 17: 113-16.

[7] Cardinalli-Neto A, Greco OT, Bestetti RB. Automatic implantable cardioverter defibrillators in Chagas' heart disease patients with malignant ventricular arrhythmias. PACE 2006; 29:467-70.

[8] Mota EA, Guimarães AC, Santana OO, Sherlock I, Hoff R, Weller TH. A nine year prospective study of Chagas disease in a defined rural population in northeast Brazil. Am J Trop Med Hyg 1990; 42: 429-40.

[9] Leveque A, De Muynck A. Chronic Chagas cardiomyopathy: methods for identifying groups at risk and/or risk behaviors. *Med Trop (Mars)* 1993; 53(4):443-53.

[10] Pazin-Filho A, Almeida-Filho OC, Furuta MS. Minor segmental wall motion abnormalities detected in patients with Chagas'disease have adverse prognostic implications..*Brazilian Journal of Medical and Biological Research* (2006) 39: 483-487).

[11] Diretrizes éticas internacionais para pesquisas biomédicas envolvendo seres humanos (CIOMS/OMS). Bioética 3: 95-134, 1995.

[12] Rose, G., Blackburn, H., Gillium, R. F., Prineas, R. J. (1982). Cardiovascular survey methods. *World Health Organization*, Monograph Series n□56.

[13] Maguire, J. H., Mott, K. E., Souza, J. A. A., Almeida, E. C., Ramos, N. B., Guimaraes, A. C. (1982). Eletrocardiographic classification and abbreviated lead system for population based studies of Chagas Disease. *Bulletin Pan American Health Organization,* 16, 47-58.

[14] Appropriateness Criteria for Transthoracic and Transesophageal Echocardiography. A Report of the American College of Cardiology Foundation Quality Strategic Directions Committee Appropriateness Criteria Working Group, American Society of Echocardiography, American College of Emergency

Physicians, American Society of Nuclear Cardiology, Society for Cardiovascular Angiography and Interventions, Society of Cardiovascular Computed Tomography, and the Society for Cardiovascular Magnetic Resonance *Endorsed by the American College of Chest Physicians and the Society of Critical Care Medicine* Douglas PS, Khandheria B, Stainback RF, Weissman NJ, Brindis RG, Patel MR, Alpert JS, Fitzgerald D, Heidenreich P, Martin ET, Messer JV, Miller AB, Picard MH, Raggi P, Reed KD, Rumsfeld JS, Steimle AE, Tonkovic R, Vijayaraghavan K, Yeon SB, Hendel RC, Peterson E, Wolk MJ, Allen JM. J Am Soc Echo. 2007; 20 (7):787-805).

[15] Garcia MJ, Thomas JD, Klein AL. New Doppler echocardiographic applications for the study of diastolic function. J Am Coll Cardiol. 1998; 32(4):865-75.

[16] Migliore RA, Guerrero FT, Armenti A, et al. Diastolic function in Chagas'Disease. Medicina.1990; 50(6):537-42.

[17] Barros MVL, Rocha MOC, Ribeiro ALP. Tissue doppler imaging in the evaluation of the regional diastolic function in Chagas' disease. Eur J Echocardiogr. 2001;2:94-99.

[18] Barros MVL, Machado FS, Ribeiro ALP. Diastolic function in Chagas' disease: An Echo and tissue doppler imaging study. Eur J Echo. 2004;5:182-88.

[19] Oliveira FP, Pedrosa RC, Gianella-Neto A. Gas exchange during exercise in different evolutional stages of chronic Chagas' heart disease. Arq Bras Cardiol 2000;75(6):481-498.

[20] Tura, BR. Aplicação do "*Data Mining*" em medicina. Tese de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 2001.

[21] Szklo M, Nieto FJ. Epidemiology beyond the basics. Maryland: Aspen Pub, 2000.

[22] Pazin-Filho A, Almeida-Filho OC, Furuta MS. Prognostic implications of minor segmental wall abnormalities in patients with Chagas' disease. J Am Coll Cardiol 1998:31 (suppl C); 339.

[23] Barros MVL, Ribeiro ALP, Machado FS, Rocha MOC. Doppler Tissular na Avaliação da Função Sistólica na Doença de Chagas. Arq Bras Cardiol. 2003; 80 (1), 31-35.

[24] Manzullo EC, Darraidon M, Libonatti O, Rozlosmk J, Chuit R. Estudio longitudinal de la cardiopatia chagásica cronica. Centro de Chagas de la Catedra de Enfermedades Infecciosas de la Facultad de Ciencias Medicas de Buenos Aires. 1982.

[25] Sternick EB, Martinelli M, Sampaio RC, Gerken LM, Teixeira RA, Scarpelli R, Scanavacca M, Nishioka SDO, Sosa, E. Sudden death in Patients with Chagas Heart Disease and Preserved Left Ventricular Function. J Cardio Electroph 2006; 17 (1), 113-116.

[26] Ferreira H. O tratamento da forma indeterminada da doença de Chagas com nifurtimox e benzonidazol. Rev Soc Bras Med Trop 1990; 23:209-211.

[27] Viotti R, Vigliano C, Armenti H, Segura E. Treatment of Chronic Chagas' disease with benzonidazole: clinical and serologic evolution of patients with long-term follow-up. A H J 127:151-162, 1994.

Apêndice G

Uso Não-convencional das Curvas de Sobrevida

O uso não-convencional das curvas de sobrevida (Seção 2.2) e dos modelos de regressão de Poisson, (Seção 2.1.4) para comparação de dados qualitativos ordinais, foi apresentado através do artigo intitulado "A non conventional use of survival curves to identify factors for gustatory alterations in patients with chronic otitis media". Os testes Log-rank, Tarone-Ware e Peto-Prentice (Seção 2.2.3) foram utilizados para a comparação das curvas de sobrevida.

O artigo foi enviado, em julho de 2010, para publicação na "Revista Brasileira de Biometria" e encontra-se sob a análise dos revisores.

A NON CONVENTIONAL USE OF SURVIVAL CURVES TO IDENTIFY FACTORS FOR GUSTATORY ALTERATIONS IN PATIENTS WITH CHRONIC OTITIS MEDIA

Basilio de Bragança Pereira PEREIRA [1] [2]
Emília Matos do NASCIMENTO [2]
Felippe FELIX [1]
Shiro TOMITA [1]

- *ABSTRACT: We present an application of discrete survival-analysis and Poisson regression to identify factors that cause gustatory alterations in patients with chronic otitis media. A case study related to a prospective study to identify factors of gustatory alterations due to chorda tympani nerve involvement in patients with chronic otitis media without prior surgery was presented. The Log-rank, Tarone-Ware, and Peto-Prentice tests pointed out a significant association between survival curves of the healthy side and the affected side of the face of patients with chronic otitis media related to gustatory alterations* ($p < 0.05$). *Significant association was also found the between the survival curves of smokers and nonsmokers patients considering the healthy side of the face* ($p < 0.05$). *The most relevant covariates identified by Poisson regression model were the side of the face, age, gender, smoke, and cholesteatoma. The proposed method can serve as an alternative procedure to statistical test for comparison of samples of discrete variables. This approach has the advantage of being more familiar to clinical researchers.*

- *KEYWORDS: Survival analysis; Kaplan-Meier; Log-rank test; Tarone-Ware test; Peto and Prentice test.*

1 Introduction

The survival time estimates can be made through parametric models, which one assumes that data may be modelled by a particular probability distribution,

[1]Federal University of Rio de Janeiro, School of Medicine and HUCFF - University Hospital Clementino Fraga Filho, P.O. Box 68507, 21941-972, Rio de Janeiro, Brazil, E-mail: *basilio@hucff.ufrj.br / felfelix@gmail.com / shiro@openlink.com.br*

[2]Federal University of Rio de Janeiro, COPPE - Postgraduate School of Engineering, P.O. Box 68507, 21941-972, Rio de Janeiro, Brazil, E-mail: *emilia@pep.ufrj.br*

or by non-parametric models. In the latter case, the Kaplan-Meier estimator is commonly used.

The product-limit estimator of Kaplan-Meier, or simply, the Kaplan-Meier estimator (Kaplan & Meier, 1958), is a estimated survival time considering that the probability that an individual will survive until a time t is independent of the probability of survival until each of the previous times. Therefore, some important concepts can be defined:
i) the probability of a death occurrence in a time interval between t_j and t_{j-1}, $j = 1, 2, \ldots, k$, given that the individual has survived beyond the immediately previous time:

$$q_j = P\big(T \in [t_{j-1}, t_j) | T \geq t_{j-1}\big)$$

ii) the probability of survival until a time t_j:

$$S(t_j) = (1 - q_1) \cdot (1 - q_2) \cdot \ldots \cdot (1 - q_j)$$

iii) and, finally, the Kaplan-Meier estimator:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right)$$

where n_j is the number of individuals at risk and d_j is the number of failures at time t_j.

In this paper we performed survival analysis method for discrete uncensored data. We used the number of taste strips instead of the usual time to event setting.

2 Data base and methods

This section presents a case study to illustrate the approach by using survival analysis to compare discrete data.

The case study is related to 45 patients with chronic otitis media (COM), being 25 with cholesteatoma and 20 without cholesteatomatous COM, with a mean age of 38 years. Eight cases of unilateral ageusia were found on the affected side. A prospective study was performed to identify gustatory alterations due to chorda tympani nerve involvement in patients with chronic otitis media (COM) without prior surgery, and to find out whether the presence of cholesteatoma worsened gustatory sensitivity in these patients.

The test was performed in patients with unilateral cholesteatomatous or suppurated COM not previously submitted to otological surgery. The test was based on "taste strips" with four different concentrations of salt, sweet, bitter, and sour. The analysis compares the two sides of the same patient, using the otological disease-free side as the control. The score could be between 0 (worst) and 16 (best), according to the number of taste strips which flavors have been recognized. The data were collected by interview and physical exam. The variables considered in this study were: age, gender, smoke, cholesteatoma, otorrhea, diabetes, hypertension,

and side of the face. The number of taste strips recognized by the patients was outcome. Patients with cholesteatomatous or suppurated COM may present gustatory alterations, even in the absence of complaints.

3 Results and discussion

A typical result of the survival curve is shown in Figure 1 for comparison of the healthy (otological disease-free side) and affected side.

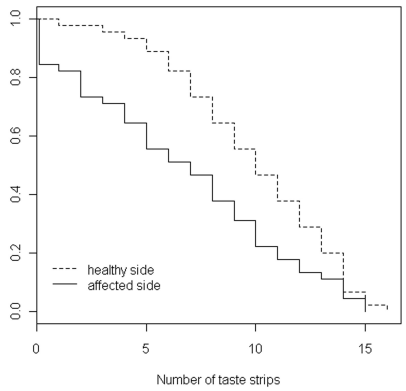


Figure - 1: Proportion of patients who recognized the flavors

Three statistical tests were performed to assess the difference between the survival curves. The results pointed out significant association between the healthy and affected side: Log-rank test ($p = 0.00908$) (Cox, 1972; Mantel, 1966; Peto & Peto, 1972), Tarone-Ware test (Tarone & Ware, 1977) ($p = 0.00239$), and Peto and Prentice test (Peto & Peto, 1972; Prentice, 1978) ($p = 0.000943$).

Survival curves were also performed for comparison of the healthy and affected side using the remaining covariates: Gender and Smoke (Figure 2); Cholesteatoma and Otorrhea (Figure 3); and Diabetes and Hypertension (Figure 4). The survivorship functions cross indicating that the proportional hazards assumption was violated.

Table 1 shows the p-values of the statistical tests to assess difference between the survival curves regarding the covariates. The only significant association was found between the survival curves of smokers and non-smokers considering the healthy side of the face.

The factors which most influences the outcome is of great interest in epidemiologic research. Poisson Regression is a class of Generalized Linear Models (Mccullagh & Nelder, 1989) often used to model count data.

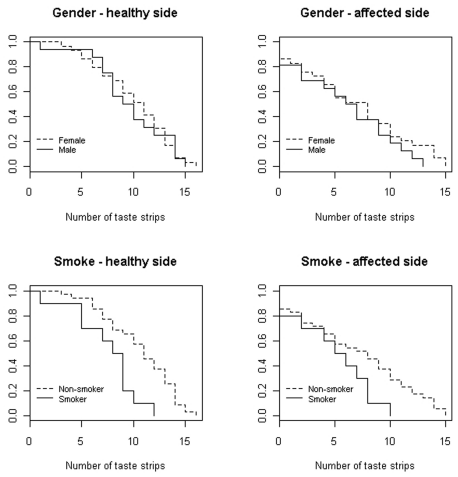


Figure - 2: Proportion of patients considering Gender and Smoke

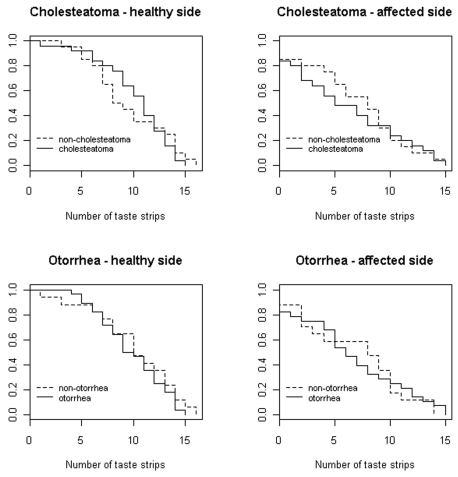


Figure - 3: Proportion of patients considering Cholesteatoma and Otorrhea

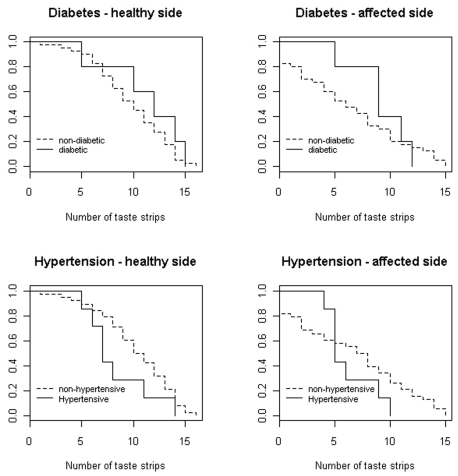


Figure - 4: Proportion of patients considering Diabetes and Hypertension

Table - 1: Statistical tests

$Covariate(side)$	$Tarone-Ware$	$Log-rank$	$Peto-Prentice$
$Gender\ (healthy)$	0.773	0.79	0.757
$Gender\ (affected)$	0.441	0.282	0.558
$Smoke\ (healthy)$	0.005	0.003	0.008
$Smoke\ (affected)$	0.115	0.064	0.187
$Cholest.\ (healthy)$	0.676	0.962	0.467
$Cholest.\ (affected)$	0.613	0.759	0.509
$Otorrhea\ (healthy)$	0.549	0.423	0.649
$Otorrhea\ (affected)$	0.895	0.937	0.806
$Diabetes\ (healthy)$	0.343	0.364	0.377
$Diabetes\ (affected)$	0.327	0.600	0.205
$Hypert.\ (healthy)$	0.107	0.136	0.097
$Hypert.\ (affected)$	0.591	0.349	0.865

Let Y_j be the number of taste strips recognized by patients with chronic otitis media and $X_1, X_2, \ldots, X_k$ a set of covariates. The Poisson regression model to this data set is

$$E(Y_j) = \lambda_j = \lambda_0\, exp(\beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj})$$

where $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the independent variables and λ_0 is the baseline hazard rate assuming that all values of the *jth* patient is zero. Therefore the Poisson regression model may be re-written as follow

$$log_e(\lambda_j/\lambda_0) = \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj}$$

where $log_e(\lambda_j/\lambda_0)$ is the log rate ratio of the *jth* patient compared to an individual with baseline characteristics.

In search for the model which best fit the data, the smallest AIC (Akaike, 1974) was the used criteria to choose the model. The best model was reached through a Poisson regression model, as shown in Table 2, with the following covariates: side of the face, age, gender, smoke, and cholesteatoma. This model provided $AIC = 538.2$.

Table - 2: Poisson model

	Estimate	*Std. Error*	*z value*	$Pr(>\|z\|)$
$(Intercept)$	3.087648	0.132424	23.316	$< 2e-16 * **$
side	−0.396536	0.074669	−5.311	$1.09e-07 * **$
age	−0.015697	0.002726	−5.758	$8.53e-09 * **$
gender	−0.253850	0.083129	−3.054	$0.00226 * *$
smoke	−0.205866	0.102337	−2.012	$0.04426*$
cholesteatoma	−0.168641	0.077527	−2.175	$0.02961*$

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusions

We believe that the proposed survival curve approach can serve as an alternative procedure to statistical test for comparison of samples from quantitative variables. In addition, this methodology may be more familiar to medical researchers.

PEREIRA, B. B; NASCIMENTO, E. M.; FELIX, F.; TOMITA, S. Uso não convencional de curvas de sobrevida para identificar os fatores de alterações gustatórias em pacientes com otite médica crônica. *Rev. Mat. Estat.,* São Paulo, v.xx, n.x, p.xx-xx, 2010. *Rev. Mat. Estat.* (São Paulo), v. xx, n.x, p. xx-xx, 2010.

■ RESUMO: Apresentamos uma aplicação de análise de sobrevida para dados discretos e regressão de Poisson para identificar os fatores que causam alterações gustatórias em pacientes com otite média crônica. Um estudo de caso envolvendo um estudo prospectivo para identificar fatores de alteração gustatória, devido ao envolvimento do nervo corda do tímpano em pacientes com otite média crônica sem cirurgia prévia foi apresentado. Os testes de Log-rank, Tarone-Ware, e Prentice-Peto apontaram uma associação significativa entre as curvas de sobrevida do lado saudável e do lado afetado da face de pacientes com otite média crônica relacionada a alterações gustatórias ($p < 0,05$). Uma associação significativa foi também encontrada entre as curvas de sobrevivida pacientes de fumantes e de não fumantes, considerando o lado saudável da face ($p < 0,05$). As covariáveis mais relevantes identificadas pelo modelo de regressão de Poisson foram lado do rosto, idade, sexo, fumo, e colesteatoma. O método proposto pode servir como um procedimento estatístico alternativo para comparação de amostras de variáveis discretas. Esta abordagem tem a vantagem de ser mais familiar aos pesquisadores clínicos.

■ PALAVRAS-CHAVE: Análise de sobrevida; Kaplan-Meier; Teste de log-rank; Teste de Tarone-Ware; Teste de Peto e Prentice.

References

AKAIKE H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v.19, p.716-723, 1974.

COX, D. R. Regression models and life tables. *Journal of the Royal Statistical Society*, Series B, v.34, p.187-220, 1972.

KAPLAN, E. L.; MEIER P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.*, v.53, p.457-481, 1958.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, v.50, p.163-170, 1966.

McCULLACH, P.; NELDER, J. A. *Generalized linear models*. 2.ed. London: Chapman and Hall, 1989. 511p.

PETO, R; PETO J. Asymptotically Efficient Rank Invariant Test Procedures (with Discussion). *Journal of the Royal Statistical Society*, Series A, v.135, p.185-207, 1972.

PRENTICE, R. L. Linear Rank Tests with Right-censored Data. *Biometrika*, v.65, p.167-179, 1978.

TARONE, R; WARE, J. On distribution-free tests for equality of survival distributions. *Biometrika*, v.64, p.156-160, 1977.

Recebido em xx.07.2010.

Aprovado após revisão em xx.xx.2010.

Apêndice H

Árvores de Regressão e Curvas de Sobrevida

As árvores de regressão (Seção 2.3.2) e as curvas de sobrevida, para análise de concordância entre medidas (Survival Agreement Plot), foram apresentadas no artigo, que tem como título, "Iodine-123-metaiodobenzylguanidine cardiac imaging as a method to detect early sympathetic neuronal dysfunction in chagasic patients with normal or borderline ECG and preserved ventricular function". O teste de Tarone-Ware (Seção 2.2.3) foi utilizado para estabelecer a comparação entre as curvas de sobrevida.

O artigo foi submetido para publicação, em junho de 2009, no "Journal of Nuclear Cardiology" e aguarda o parecer dos revisores.

Cover Page

Iodine-123-metaiodobenzylguanidine cardiac imaging as a method to detect early sympathetic neuronal dysfunction in chagasic patients with normal or borderline ECG and preserved ventricular function

Journal:	*Journal of Nuclear Cardiology*
Manuscript ID:	JNC-09-105-OA
Manuscript Type:	Original Article
Date Submitted by the Author:	25-Jun-2009
Complete List of Authors:	Landesmann, Maria Carolina; Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Department of Radiology Barbosa da Fonseca, Lea; Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Department of Radiology Pereira, Basilio; Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Department of Radiology; Programa de Engenharia de Produção, COPPE, Universidade Federal do Rio de Janeiro Nascimento, Emilia; Programa de Engenharia de Produção, COPPE, Universidade Federal do Rio de Janeiro Souza, Sergio Augusto; Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Department of Radiology Rosado de Castro, Paulo Henrique; Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Department of Radiology Pedrosa, Roberto; Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Department of Cardiology; Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Department of Radiology
Keywords:	MIBG imaging, Myocarditis, SPECT, Cardiomyopathy, Inflammation

Title Page

Iodine-123-metaiodobenzylguanidine cardiac imaging as a method to detect early sympathetic neuronal dysfunction in chagasic patients with normal or borderline ECG and preserved ventricular function.

Running head: 123I-MIBG imaging in asymptomatic chagas disease

Maria Carolina P. Landesmann[a,c], Lea Mirian Barbosa da Fonseca[a,c], Basilio de B. Pereira[a,b], Emília M. do Nascimento[b], Sergio Augusto Lopes de Souza[a], Paulo Henrique Rosado de Castro[a], Roberto C. Pedrosa[a]

[a]Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

[b]Programa de Engenharia de Produção, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

[c]Proecho/Hospital Samaritano, Rio de Janeiro, Brazil

Authors:

Maria Carolina P. Landesmann, MD, MSc

Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Author contributions: Contributed to the conception and design, analysis and interpretation of data; drafting of the manuscript and final approval of the manuscript submitted.

Lea Mirian Barbosa da Fonseca, MD, PhD

Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Author contributions: Contributed to the conception and design, analysis and interpretation of data; drafting of the manuscript and final approval of the manuscript submitted.

Basilio de Bragança Pereira, PhD

Hospital Universitário Clementino Fraga Filho and Programa de Engenharia de Produção, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Author contributions: Contributed to the conception and design, analysis and interpretation of data; drafting of the manuscript and final approval of the manuscript submitted.

Emília Matos do Nascimento, MSc

Programa de Engenharia de Produção, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Author contributions: Contributed to the analysis and interpretation of data; drafting of the manuscript and final approval of the manuscript submitted.

Sergio Augusto Lopes de Souza, MSc, PhD

Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Author contributions: Contributed to the analysis and interpretation of data; drafting of the manuscript and final approval of the manuscript submitted.

Paulo Henrique Rosado de Castro

Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Author contributions: Contributed to the analysis and interpretation of data; drafting of the manuscript and final approval of the manuscript submitted.

Roberto Coury Pedrosa, MD, PhD

Hospital Universitário Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Author contributions: Contributed to the conception and design, analysis and interpretation of data; drafting of the manuscript and final approval of the manuscript submitted.

Corresponding author address:

Lea Mirian Barbosa da Fonseca, MD, PhD, Hospital Universitário Clementino Fraga Filho, Departamento de Radiologia, subsolo, Universidade Federal do Rio de Janeiro, Ilha do Fundão, 21949-913, Rio de Janeiro, Brazil. Telephone: 55-21-2562-2399; Fax: 55-21-2562-2399; e-mail: lmirian@hucff.ufrj.br

Abstract (200 words)

Background: The indeterminate form of Chagas disease represents the most common chronic presentation. The aim of this study was to assess cardiovascular autonomic system function with ^{123}I-MIBG scintigraphy in chagasic patients with normal or “borderline” electrocardiographic alterations and preserved left ventricular function.

Methods: A total of 40 chagasic patients and 19 control subjects were included in the study. Patients had normal echocardiogram and chest radiography, no arrhythmias or myocardial ischemia and normal exercise performance for age, gender and body mass index. ^{123}I-MIBG scintigraphy was performed and the heart-to-mediastinum (H/M) uptake was used as the primary predictor in the present analysis. The data analysis was performed by using Non-parametric Regression Trees and the Survival Agreement Plot.

Results: Variables analyzed in the regression tree were age, sex, 20 minutes and 3h H/M uptake after injection of ^{123}I-MIBG, washout rate and SPECT imaging. The 3h H/M ratio was the only significant variable ($p < 0.001$) and for 95% of chagasic patients this value was below 2.19.

Conclusions: This study presents evidence that, even in chagasic subjects with normal or “borderline” electrocardiogram with preserved ventricular function, cardiac autonomic sympathetic modulation is affected.

Keywords: Chagas' heart disease; Cardiac sympathetic nervous activity; Iodine-123 metaiodobenzylguanidine; electrocardiogram; preserved ventricular function;

Introduction

Chagas' disease, caused by the protozoan *Trypanosoma cruzi* and transmitted by triatomid bugs, is characterized by several clinical disease forms. It is one of the major public health problems in almost all Latin and Central American countries. Also, Chagas´ disease may represent a noteworthy healthcare-related problem in Europe and the United States due to elevated immigration index (1). Recent estimates from the World Health Organization indicate that 18 million persons are chronically infected and 200.000 new cases occur each year (1).

The disease has acute and chronic phases characterized by distinct signs and symptoms, separated by an indeterminate period, which is characterized by a discrete to moderate focal or zonal myocarditis and/or neuroganglionitis that may persist and last for years without clinical signs of cumulative damage (2). This period represents the most common chronic form of the disease, accounting for about 70% of the cases, usually lasts for 10 to 30 years, but in a majority it can persists throughout life, and still has an intriguing and incompletely known pathophysiology and course (3). Nevertheless, this form may be considered as belonging to the category of potential cardiac patients, since each year approximately 3% of them will develop involvement of the heart (4).

There is not enough evidence to support the need for immediate treatment in all patients who present with indeterminate form of Chagas' disease, although there are studies suggesting an increased cardiovascular morbidity and mortality in patients only with "borderline" electrocardiographic alterations (5, 6). It is well

known that those who suffer from chronic heart failure are at higher risk for sudden death (7, 8). Patients with heart failure are usually under medical care, and risk stratification for primary prevention of sudden death is a feasible task. On the other hand, there is no consensus on the best strategy to carry out risk stratification for primary prophylaxis of sudden death in chagasic patients without apparent heart disease with New York Heart Association (NYHA) functional class I symptoms. Cardiovascular autonomic disturbance is a mortality predictor in post acute myocardial infarction patients (9), in those who suffer from cardiac heart failure (10) and in diabetes mellitus (11).

Studies with iodine-123 metaiodobenzylguanidine (^{123}I-MIBG) scintigraphy have been established as a reliable and reproducible marker of cardiovascular autonomic function (12). ^{123}I-MIBG scintigraphy has also been proposed as a method to evaluate sympathetic cardiac nerve damage that occurred at the ventricular myocardial level (13).

The aim of this study was to assess cardiovascular autonomic system function with ^{123}I-MIBG scintigraphy in chagasic patients with normal or “borderline” electrocardiographic alterations and preserved left ventricular function

Methods

Patient Population

This is a cross-sectional study carried out from October 2003 to August 2007. The selection of patients was carried out at the Chagas’ disease outpatient clinic from the University Hospital of our institution (n=76; 48 patients with normal

electrocardiogram and 28 with “borderline” electrocardiogram). All patients have been regularly followed up by several years by means of periodical clinical and laboratory examination since Chagas’disease diagnosis was made. Patients with any chronic illness (specially cardiovascular, including hypertension, diabetes mellitus); patients with left ejection fraction lower than 55%; patients with thyroid or renal dysfunction, with chronic obstructive pulmonary disease, neuropathy; patients taking drugs that could interfere with the cardiovascular system; pregnancy; alcoholism, smoking, and those using steroidal drugs,; cardiac arrhythmias, other cardiomyopathy, prior coronary disease and Parkinson's disease were all excluded. The remaining 40 patients (22 men and 18 women; age range, 29 to 78 yr, mean age 56.3 yr) with a definite serologic status for Chagas’ disease (2 different positive reactions to T cruzi) had a normal clinical history and physical examination and the following normal exams: full blood count, free T4, thyroid-stimulating hormone, glucose, potassium, creatinine, blood urea nitrogen. All patients had a 24h Holter register free from arrhythmias, a normal echocardiogram, a normal chest radiography and absence of myocardial ischemia and a normal exercise performance for age, gender and body mass index in the treadmill exercise test. Nineteen healthy volunteers (7 men and 12 women; age range, 23 to 72 yr; mean age 42.6 yr) were used as control subjects to determine reference ranges of ^{123}I-MIBG activity and washout kinetics. All control and chagasic subjects examined were in excellent physical and mental conditions, in regular daily activities and not using any drug. The body mass index was below 30 Kg/m^2 for all subjects and similar (p=0.41) for the

chagasic (21.5±1 Kg/m^2) and control groups (24.2 +/-1.5 Kg/m^2). Chagasic patients did not receive anti-T cruzi specific drug treatment. No clinical problems occurred during the study. Both groups gave written informed consent for their participation in the study, which was approved by the local review committee.

Electrocardiographic analysis

Tracing were coded independently by 2 cardiologists using a modification of the Minnesota Code (14). If there was a disagreement in coding, the tracing was reviewed by both, and a consensus reached. The criteria used were:

Normal. Absence of any alteration listed under “borderline” with normal sinus rhythm. The following were allowable: tall R wave in V5 if not associated with T wave or ST segment alterations and rr’ in V1 if other criteria for incomplete right bundle branch block were not met.

Borderline. Absence of any alteration listed under “normal” plus one or more of the following: minimal Q wave alterations; right axis deviation if other criteria for complete right bundle branch block were not met; tall precordial R waves without T wave or segment alterations; minor T wave or ST segment alterations; sinus arrhythmia or sinus tachycardia; sinus bradycardia; non-repetitive atrial or junctional extra-systoles; unifocal non-repetitive ventricular junctional extrasystoles; atrioventricular junctional rhythm and wide P wave if associated with sinus tachycardia.

^{123}I-MIBG Imaging

To avoid excessive irradiation in the thyroid gland, oral administration of potassium iodine syrups was recommended the day before the investigation until the next day. Patients were also instructed to remain fasted for 4 hours before the study. Seven mCi (259 MBq) of ^{123}I-MIBG were injected intravenously in 1 minute through an indwelling catheter. Planar images were acquired in anterior and 45^{o} left anterior oblique views of the thorax using a dual head gamma camera equipped with two high resolution collimators (Millennium MG General Electric Medical Systems, Milwaukee, WI) (15, 16). Images were recorded 20 minutes and 3h after injection with an acquisition time of 5 minutes, matrix size 256x256, and zoom factor x 1.6 (17). Energy discrimination was achieved by a 20% energy window centered on the 159 keV photo peak of Iodine-123. After the planar scan, SPECT imaging was performed between 60 and 90 minutes after the injection, and data were acquired from a 45^{o} right anterior oblique angle over a 180^{o} arc in 36 preset sampling angles for 30s per angle (18). The data were stored in a 64x64 matrix. The early and late planar images were analyzed qualitatively and semi-quantitatively by two observers.

A heart-to-mediastinum ratio (H/M ratio) of ^{123}I-MIBG uptake was used as the primary predictor in the present analysis. ^{123}I-MIBG uptake is semiquantified by calculating a H/M ratio after drawing regions of interest over the upper mediastinum in the anterior view and over the myocardium in the left anterior oblique view. The H/M ratio was calculated without background subtraction as average counts per pixel in the myocardium divided by average counts per pixel in the mediastinum. The clearance rate from myocardium (washout rate) was

calculated by: (Initial myocardial ^{123}I-MIBG uptake – Delayed myocardial ^{123}I-MIBG uptake/ Initial ^{123}I-MIBG uptake) x 100. The tomographic images were processed through the QGS / QPS protocol and the data were reconstructed in short-axis, horizontal long-axis, and vertical long-axis views. Visual evaluation was performed and changes of segmental uptake of ^{123}I-MIBG in the heart were described.

Statistical Analysis

The data analysis was performed by using Non-parametric Regression Trees and the Survival Agreement Plot. Non-parametric regression tree is a recursive partitioning method based on decision rules. The recursive partitioning computational algorithm was developed by Hothorn et al. (19) and implemented into a well defined theory of conditional inference procedures, with unbiased variable selection and statistical stopping criterion. Each node of the tree presents the p-value which corresponds to the log-rank test. Afterwards, the logistic regression was implemented to select the most significant variables and to confirm the results obtained through the tree based method.

The interobserver and intraobserver agreement analysis was performed by using the Survival Agreement Plot which was proposed by Luiz et al. (20) to assess the reliability of a quantitative measure. This method implements the Kaplan Meier curves without censored data where the failures occur at the absolute difference between the observer scores. An improved method proposed by Llorca and Delgado-Rodríguez (21) was also used. This method considers two groups of the

real score differences instead of the global differences. The equality of the two survivor functions obtained through the Llorca and Delgado-Rodriguez method was evaluated by using the Tarone-Ware test (22), which is a non-parametric weighted rank test. The R software (23) was used to implement all the data analysis.

Results

The general characteristics of patients are presented in tables 1 and 2.The final tree of the analysis is given in figure 1. The variables in this tree are age, sex, 20 min and 3h H/M ratio, washout rate and SPECT imaging. The 3h H/M ratio was the only significant variable ($p < 0.001$). The value of 2.19 had a sensitivity of 95% and a specificity of 57.9%. The positive predictive value was 0.82 and the negative predictive value was 0.84 indicating an acceptable model.

For interobserver and intraobserver agreements, the survival-agreement plot proposed by Luiz et al shows that a tolerance limit greater than 0.3 and 0.15, respectively, in the gauge of the continuous measures for the 3h H/M ratio, would generate a proportion of disagreement near 0% and 10%, respectively (figures 2 and 3). The survival-agreement plot using the Llorca et al approach, in the case group, shows that the disagreement proportions were not significant ($p= 0.07$ and $p=0.70$ for interobserver and intraobserver agreements, respectively) (figures 2 and 3). In the control group, the disagreement proportions were not statistically different for intraobserver agreement (0.08). For the interobserver agreement

(p=0.03), the disagreement proportions were statistically different, but was clinically insignificant (threshold < 0.3) (figures 2 and 3).

In the chagasic group, 25 of 40 (62.5%) had “borderline” ECG, while all control patients had normal ECG. In the analysis of 3h H/M ratio there was a significant difference between chagasic patients with normal ECG, chagasic patients with “borderline” ECG and control subjects, what is demonstrated in figure 4. An example of 3h ^{123}I-MIBG planar images of a control subject and a chagasic patient are illustrated in figure 5. SPECT images of these same patients are illustrated in figure 6.

Discussion

Since the introduction of ^{123}I-MIBG for cardiac imaging in 1981, the assessment of pre-synaptic sympathetic function with the analogue of norepinephrine has been widely used in Europe and Asia for the study of various cardiac conditions. However, information obtained using this radiotracer are not yet fully understood, especially in some cardiomyopathies such as Chagas heart disease (24).

A recent report analyzed ^{123}I-MIBG scintigraphies of 37 patients in various stages of Chagas’ cardiomyopathy and found that in asymptomatic subjects with no cardiac involvement a considerable proportion of patients (4 of 12; 33%) had significant reduction in regional ^{123}I-MIBG uptake. This study pointed toward regional sympathetic denervation at the ventricular myocardial level (13).

One possible mechanism for the pathogenesis of cardiac autonomic sympathetic involvement is the chronic inflammatory damage of peripheral ganglia and/or neuronal fibers and perhaps of central autonomic structures, with or without neuroganglionic depopulation in different intensity or stages of evolution (2, 25). In the indeterminate form, a sustained involvement of the cardiac intrinsic autonomic innervation (26)(2, 27) and neurotransmitter receptors at the cellular membrane (28) by infiltrating T cells in myocardial fibers is thought to be a factor contributing to autonomic dysfunction. Moreover, it has been argued that the fixation of cardiac neurotransmitter receptors by anti-receptor antibodies contributes to the autonomic dysfunction (28-32). Although a strong association between circulating anti-receptor antibodies and cardiac autonomic dysfunction has been verified (33), it is not known whether the presence and/or titer of anti-receptor antibodies correlates with cardiac autonomic dysfunction.

The clinical significance of the cardiac autonomic dysfunction in some chagasic patients with the indeterminate form is a more difficult question, considering the good long-term prognosis of this form of the disease (2, 27). A recent review has suggested that for asymptomatic patients with nonspecific ECG changes (eg, rsR' not meeting criteria for right bundle-branch block or a minor increase in PR interval), the need for further evaluation should be judged on an individual basis (34). Impaired autonomic control of heart may be a contributing factor for electrical cardiac disturbances so common in Chagas'disease (35). According to Myerburg et al (36) the occurrence of complex, cardiac arrhythmias depends on (1) the existence of a substratus - in

Chagas´ heart disease this can be represented by areas of inflammation and necrosis causes by initial acute myocarditis; (2) triggering factors - chagasic patients have premature ventricular contractions as an usual manifestation of their disease; and (3) arrythmogenic factors - in Chagas disease this is represented by the unbalance between sympathetic and parasympathetic nervous system.

Classical studies by Carrió et al (16) established protocols for the evaluation of cardiac innervation using ^{123}I-MIBG scintigraphy and oriented the present study. We used a similar methodology through the selection of regions of interest in the mediastinum and in the heart in planar images in anterior and left anterior oblique views (the latter is used to minimize the superposition of heart and lung), respectively. The H/M ratio was determined through the average counts per pixel. However, parameters such as the H/M ratio and washout have variable reference values in the literature, as described in a recent review by Patel and Iskandrian (37), in which a meta-analysis of 7 studies and a total of 96 healthy subjects demonstrated that the H/M ratio varied from 2.1+/-0.2 to 2.8 +/-0.6 with a medium of 2.5 +/- 0.3 and washout values varied from 6 +/- 22 to 28 +/- 3 with a medium of 20 +/- 10.

Therefore, we decided to use a classification tree including all 40 chagasic patients and 19 control subjects to establish a cutoff value for this sample and we obtained a statistically significant value for the 3h H/M ratio of 2.19, which is in accordance to the literature. We believe that the normal values previously

determined as H/M ratio > 1.8 and the washout ≤ 10% (16) may not be adequate to our population`s characteristics.

It is valid to remember that the semi-quantitative analysis to determine these parameters has limitations such as the superposition of non-cardiac structures, such as the lung and mediastinum, by other heart segments and movement artifacts. In this sense, the use of single photon emission computed tomography (SPECT) may be useful, allowing more accurate segmental analysis.

Conclusion

This study presents evidence that, even in chagasic subjects with normal or “borderline” electrocardiogram with preserved ventricular function and New York Heart Association (NYHA) functional class I, cardiac autonomic sympathetic modulation is affected.

References

(1) World Health Organization. Control of Chagas disease : second report of the WHO expert committee. Geneva: World Health Organization; 2002.
(2) Marin-Neto JA, Cunha-Neto E, Maciel BC, Simoes MV. Pathogenesis of chronic Chagas heart disease. Circulation 2007;115:1109-23.
(3) Dias JC. The indeterminate form of human chronic Chagas' disease A clinical epidemiological review. Rev Soc Bras Med Trop 1989;22:147-56.
(4) Pinto Dias JC. The treatment of Chagas disease (South American trypanosomiasis). Ann Intern Med 2006;144:772-4.
(5) Mota EA, Guimaraes AC, Santana OO, Sherlock I, Hoff R, Weller TH. A nine year prospective study of Chagas' disease in a defined rural population in northeast Brazil. Am J Trop Med Hyg 1990;42:429-40.
(6) Rassi A, Jr., Rassi SG, Rassi A. Sudden death in Chagas' disease. Arq Bras Cardiol 2001;76:75-96.
(7) Rassi A, Jr., Rassi A, Rassi SG. Predictors of mortality in chronic Chagas disease: a systematic review of observational studies. Circulation 2007;115:1101-8.

(8) Rassi A, Jr., Rassi A, Little WC, Xavier SS, Rassi SG, Rassi AG et al. Development and validation of a risk score for predicting death in Chagas' heart disease. N Engl J Med 2006;355:799-808.
(9) Kleiger RE, Miller JP, Bigger JT, Jr., Moss AJ. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. Am J Cardiol 1987;59:256-62.
(10) Nolan J, Batin PD, Andrews R, Lindsay SJ, Brooksby P, Mullen M et al. Prospective study of heart rate variability and mortality in chronic heart failure: results of the United Kingdom heart failure evaluation and assessment of risk trial (UK-heart). Circulation 1998;98:1510-6.
(11) Malpas SC, Maling TJ. Heart-rate variability and cardiac autonomic function in diabetes. Diabetes 1990;39:1177-81.
(12) Tsuchimochi S, Tamaki N, Tadamura E, Kawamoto M, Fujita T, Yonekura Y et al. Age and gender differences in normal myocardial adrenergic neuronal function evaluated by iodine-123-MIBG imaging. J Nucl Med 1995;36:969-74.
(13) Simoes MV, Pintya AO, Bromberg-Marin G, Sarabanda AV, Antloga CM, Pazin-Filho A et al. Relation of regional sympathetic denervation and myocardial perfusion disturbance to wall motion impairment in Chagas' cardiomyopathy. Am J Cardiol 2000;86:975-81.
(14) Maguire JH, Mott KE, Souza JA, Almeida EC, Ramos NB, Guimaraes AC. Electrocardiographic classification and abbreviated lead system for population-based studies of Chagas' disease. Bull Pan Am Health Organ 1982;16:47-58.
(15) Abe M, Hamada M, Matsuoka H, Shigematsu Y, Sumimoto T, Hiwada K. Myocardial scintigraphic characteristics in patients with primary aldosteronism. Hypertension 1994;23:I164-7.
(16) Carrio I. Cardiac neurotransmission imaging. J Nucl Med 2001;42:1062-76.
(17) Lanza GA, Giordano A, Pristipino C, Calcagni ML, Meduri G, Trani C et al. Abnormal cardiac adrenergic nerve function in patients with syndrome X detected by [123I]metaiodobenzylguanidine myocardial scintigraphy. Circulation 1997;96:821-6.
(18) Wakabayashi T, Nakata T, Hashimoto A, Yuda S, Tsuchihashi K, Travin MI et al. Assessment of underlying etiology and cardiac sympathetic innervation to identify patients at high risk of cardiac death. J Nucl Med 2001;42:1757-67.
(19) Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics 2006;15:651-74.
(20) Luiz RR, Costa AJ, Kale PL, Werneck GL. Assessment of agreement of a quantitative variable: a new graphical approach. J Clin Epidemiol 2003;56:963-7.
(21) Llorca J, Delgado-Rodriguez M. Survival analytical techniques were used to assess agreement of a quantitative variable. J Clin Epidemiol 2005;58:314-5.
(22) Tarone R, Ware J. On distribution-free tests for equality of survival distributions. Biometrika 1977;64:156-60
(23) Team RDC. R: A language and environment for statistical computing. . Vienna, Austria: R Foundation for Statistical Computing; 2008.
(24) Chen GP, Tabibiazar R, Branch KR, Link JM, Caldwell JH. Cardiac receptor physiology and imaging: an update. J Nucl Cardiol 2005;12:714-30.

(25) Junqueira Junior LF, Beraldo PS, Chapadeiro E, Jesus PC. Cardiac autonomic dysfunction and neuroganglionitis in a rat model of chronic Chagas' disease. Cardiovasc Res 1992;26:324-9.
(26) Oliveira JS. A natural human model of intrinsic heart nervous system denervation: Chagas' cardiopathy. Am Heart J 1985;110:1092-8.
(27) Junqueira LF, Jr. A summary perspective on the clinical-functional significance of cardiac autonomic dysfunction in Chagas' disease. Rev Soc Bras Med Trop 2006;39 Suppl 3:64-9.
(28) Hernandez CC, Barcellos LC, Gimenez LE, Cabarcas RA, Garcia S, Pedrosa RC et al. Human chagasic IgGs bind to cardiac muscarinic receptors and impair L-type Ca2+ currents. Cardiovasc Res 2003;58:55-65.
(29) de Oliveira SF, Pedrosa RC, Nascimento JH, Campos de Carvalho AC, Masuda MO. Sera from chronic chagasic patients with complex cardiac arrhythmias depress electrogenesis and conduction in isolated rabbit hearts. Circulation 1997;96:2031-7.
(30) Masuda MO, Levin M, De Oliveira SF, Dos Santos Costa PC, Bergami PL, Dos Santos Almeida NA et al. Functionally active cardiac antibodies in chronic Chagas' disease are specifically blocked by Trypanosoma cruzi antigens. FASEB J 1998;12:1551-8.
(31) Retondaro FC, Dos Santos Costa PC, Pedrosa RC, Kurtenbach E. Presence of antibodies against the third intracellular loop of the m2 muscarinic receptor in the sera of chronic chagasic patients. FASEB J 1999;13:2015-20.
(32) Gimenez LE, Hernandez CC, Mattos EC, Brandao IT, Olivieri B, Campelo RP et al. DNA immunizations with M2 muscarinic and beta1 adrenergic receptor coding plasmids impair cardiac function in mice. J Mol Cell Cardiol 2005;38:703-14.
(33) Sterin-Borda L, Borda E. Role of neurotransmitter autoantibodies in the pathogenesis of chagasic peripheral dysautonomia. Ann N Y Acad Sci 2000;917:273-80.
(34) Bern C, Montgomery SP, Herwaldt BL, Rassi A, Jr., Marin-Neto JA, Dantas RO et al. Evaluation and treatment of chagas disease in the United States: a systematic review. JAMA 2007;298:2171-81.
(35) Medei E, Pedrosa RC, Benchimol Barbosa PR, Costa PC, Hernandez CC, Chaves EA et al. Human antibodies with muscarinic activity modulate ventricular repolarization: basis for electrical disturbance. Int J Cardiol 2007;115:373-80.
(36) Myerburg RJ, Kessler KM, Bassett AL, Castellanos A. A biological approach to sudden cardiac death: structure, function and cause. Am J Cardiol 1989;63:1512-6.
(37) Patel AD, Iskandrian AE. MIBG imaging. J Nucl Cardiol 2002;9:75-94.

Tables:

Table 1: Summary of characteristics and ^{123}I-MIBG analysis for the control group

Patient	Sex	Age	ECG	20 min H/M	3 h H/M	washout	SPECT
AJS	F	26	N	2,19	2,43	16%	N
BC	F	23	N	2,05	2,37	19%	N
CAGM	F	34	N	2,06	2,45	20%	N
CES	M	39	N	2,53	2,64	25%	N
ETO	M	71	N	1,89	1,68	35%	N
FJL	F	42	N	1,85	1,87	28%	N
FLG	M	55	N	2,28	2,22	11%	N
GCF	F	30	N	2,64	2,76	22%	N
HL	M	72	N	1,77	1,81	29%	N
JCF	M	43	N	2,17	2,28	21%	N
JO	F	45	N	2,1	2,15	21%	N
LMBF	F	59	N	2,01	2,28	30%	N
MCCP	F	36	N	2,39	2,5	30%	N
MCPP	F	34	N	2,2	2,25	20%	N
MRC	F	40	N	1,73	1,76	34%	N
MVFTO	F	26	N	1,86	1,95	19%	N
RCP	M	60	N	1,88	1,97	20%	N
RSLL	M	43	N	2,22	2,26	24%	N
SCS	F	31	N	2,1	2,13	36%	A
All control subjects		42,58±14,61		2,10±0,24	2,20±0,29	24,0±6,7%	

M = male; F = female; N = normal; B = borderline; A = altered

Table 2: Summary of characteristics and ^{123}I-MIBG analysis for the chagasic group

Patient	Sex	Age	ECG	20 min H/M	3 h H/M	washout	SPECT
AFB	M	54	N	2,07	2,02	25%	N
AMMP	F	38	N	2,05	2,19	26%	N
ASS	M	56	N	2,19	2,27	32%	N
DPA	F	33	N	1,88	2,15	23%	N
GPA	M	31	N	2,21	2,11	17%	N
GPD	M	55	N	1,8	1,68	29%	A
GS	M	67	N	1,85	1,98	23%	A
JC	M	53	N	1,97	2,06	24%	N
JR	M	72	N	2,09	1,99	27%	N
LAAM	M	65	N	1,75	1,8	24%	N
LAS	M	72	N	2,47	2,29	24%	N
LFB	F	39	N	1,79	2,15	28%	N
MCP	F	45	N	2,25	2,1	20%	N
MLMC	M	46	N	1,6	1,72	26%	A
RVF	M	51	N	2,38	2,02	31%	N
Normal ECG		51,8±13,29		2,02±0,24	2,03±0,18	25±3,9%	
AFR	M	29	B	1,8	2,04	11%	N
ANS	M	64	B	1,95	1,49	35%	A
ATS	M	42	B	1,73	1,79	25%	A
ECS	F	64	B	1,96	1,46	29%	N
EDE	M	67	B	1,99	1,88	26%	A
GGA	M	60	B	1,77	1,79	21%	N
GMA	F	76	B	1,76	1,48	29%	A
HGS	M	56	B	2,2	2,17	33%	N
JF	M	30	B	1,7	1,46	28%	A
JFS	F	60	B	1,96	2,02	29%	A
JPS	M	78	B	1,74	1,8	27%	A
JS	M	72	B	2,1	1,98	18%	N
LGR	M	41	B	2,06	2,15	29%	N
MCP	F	60	B	1,88	1,64	39%	A
MGS	F	59	B	1,85	1,93	33%	N
MJF	F	66	B	1,9	1,84	22%	N
MMPN	F	64	B	2,05	1,84	29%	N
MMPS	F	59	B	1,55	1,54	29%	A
NG	F	45	B	1,87	1,6	31%	A
NHS	F	69	B	2,06	1,68	34%	N
OSA	F	69	B	1,53	1,64	20%	N
QMC	F	55	B	1,74	1,79	26%	N
SCS	M	73	B	1,67	1,82	22%	A
SDRF	F	58	B	1,64	1,45	29%	A
VSS	F	61	B	1,52	1,09	41%	A
“Borderline” ECG		59,08±12,88		1,83±0,18	1,73±0,25	27,8±6,5%	
All chagasic patients		56,35±13,35		1,90±0,22	1,84±0,27	26,85±5,7%	
M = male; F = female; N = normal; B = borderline; A = altered							

Figure legends

Figure 1. Regression tree of the analysis of variables age, sex, 20 min and 3h H/M ratio, washout rate and SPECT images.

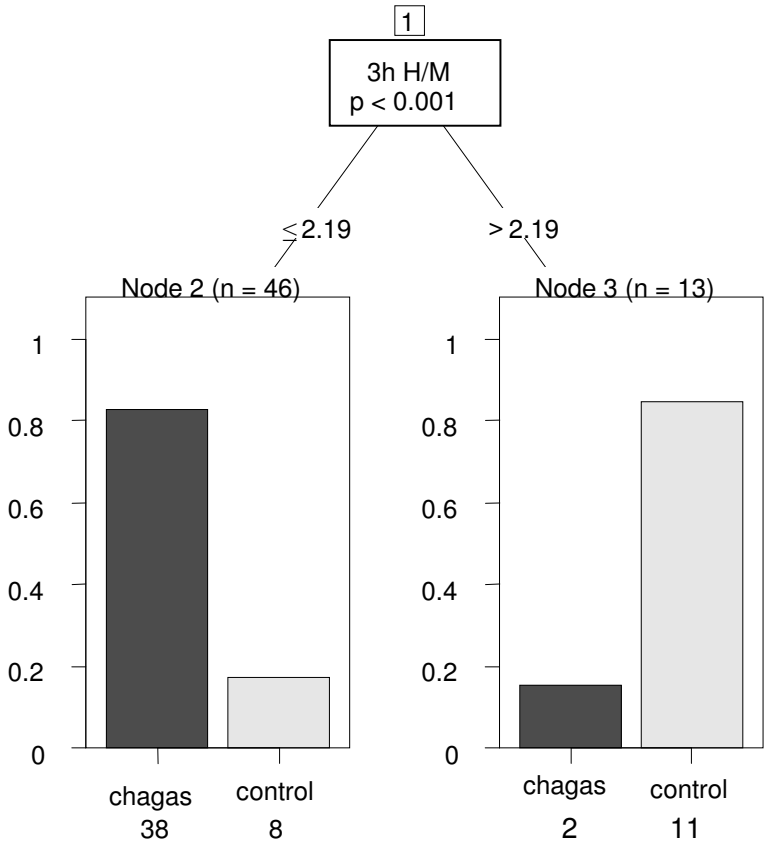
Figure 2. Analysis of disagreement proportions for intraobserver agreement by Luiz and Llorca methods.

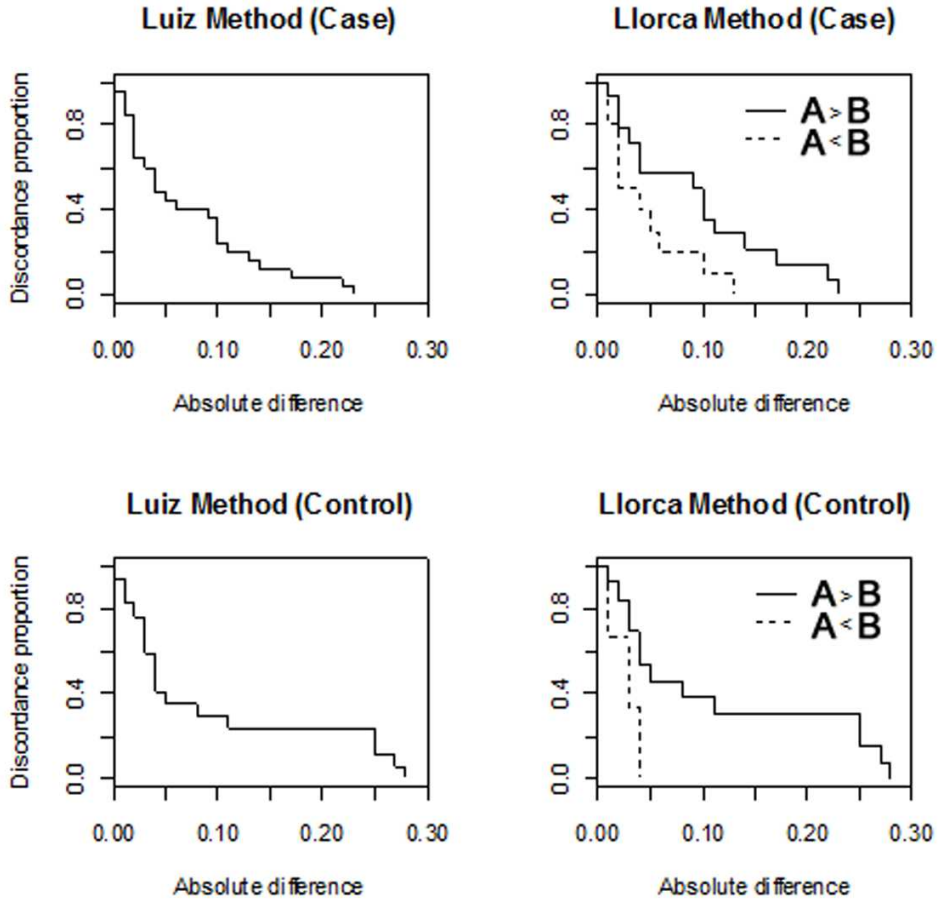
Figure 3. Analysis of disagreement proportions for interobserver agreement by Luiz and Llorca methods.

Figure 4. Boxplots demonstrating the distribution of 3h H/M ratio for chagasic patients with normal ECG, chagasic patients with "borderline" ECG and control subjects.

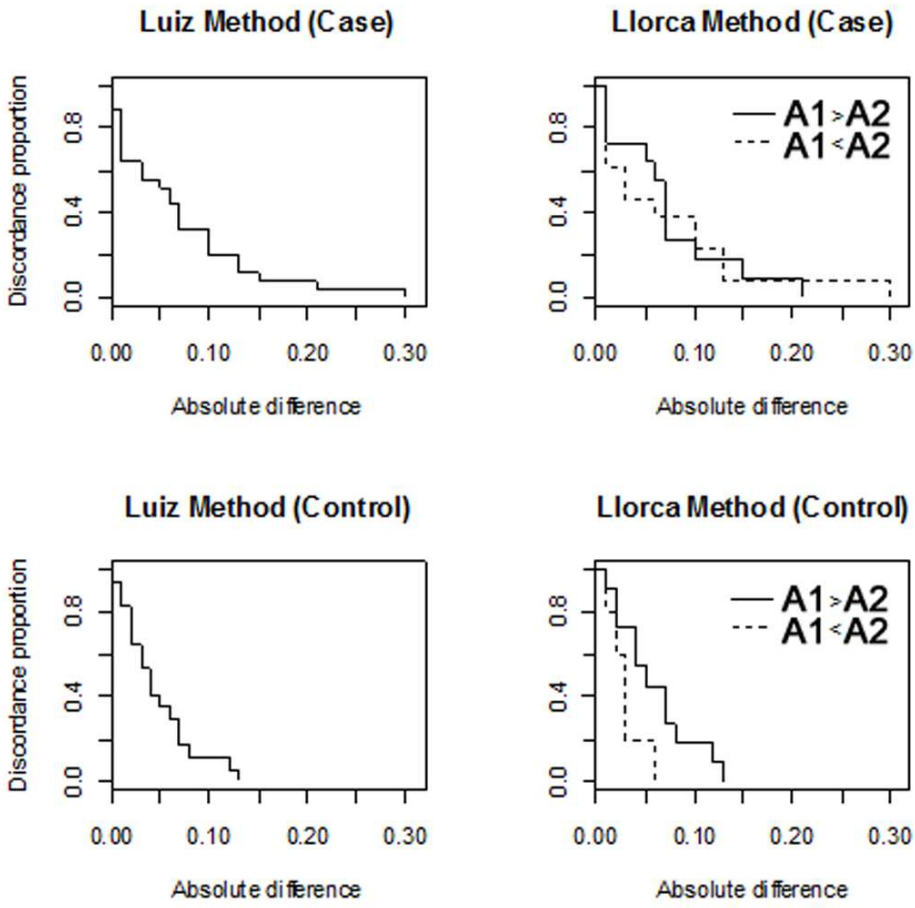
Figure 5. Planar images in anterior (**A**) and LAO projections (**B**) illustrate the normal pattern of distribution of ^{123}I-MIBG in a control subject. Planar images in anterior (**C**) and posterior (**D**) projections illustrate an altered pattern of distribution of ^{123}I-MIBG in the heart of a chagasic patient with "borderline" ECG.

Figure 6. SPECT images illustrate the normal pattern of distribution of ^{123}I-MIBG in a control subject (**A**). SPECT images illustrate a decreased uptake of ^{123}I-MIBG in apical and inferior ventricular walls in a chagasic patient with "borderline" ECG (**B**).

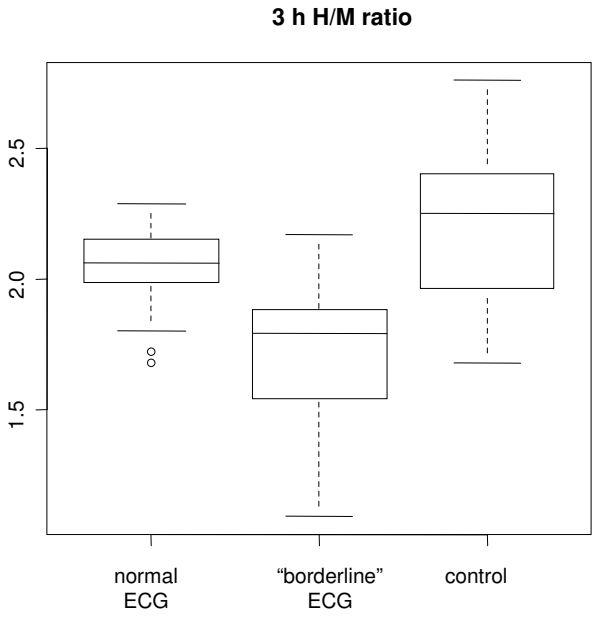

1
3h H/M
p < 0.001
≤ 2.19
> 2.19
Node 2 (n = 46)
Node 3 (n = 13)
1
0.8
0.6
0.4
0.2
0
chagas
control
38
8
2
11

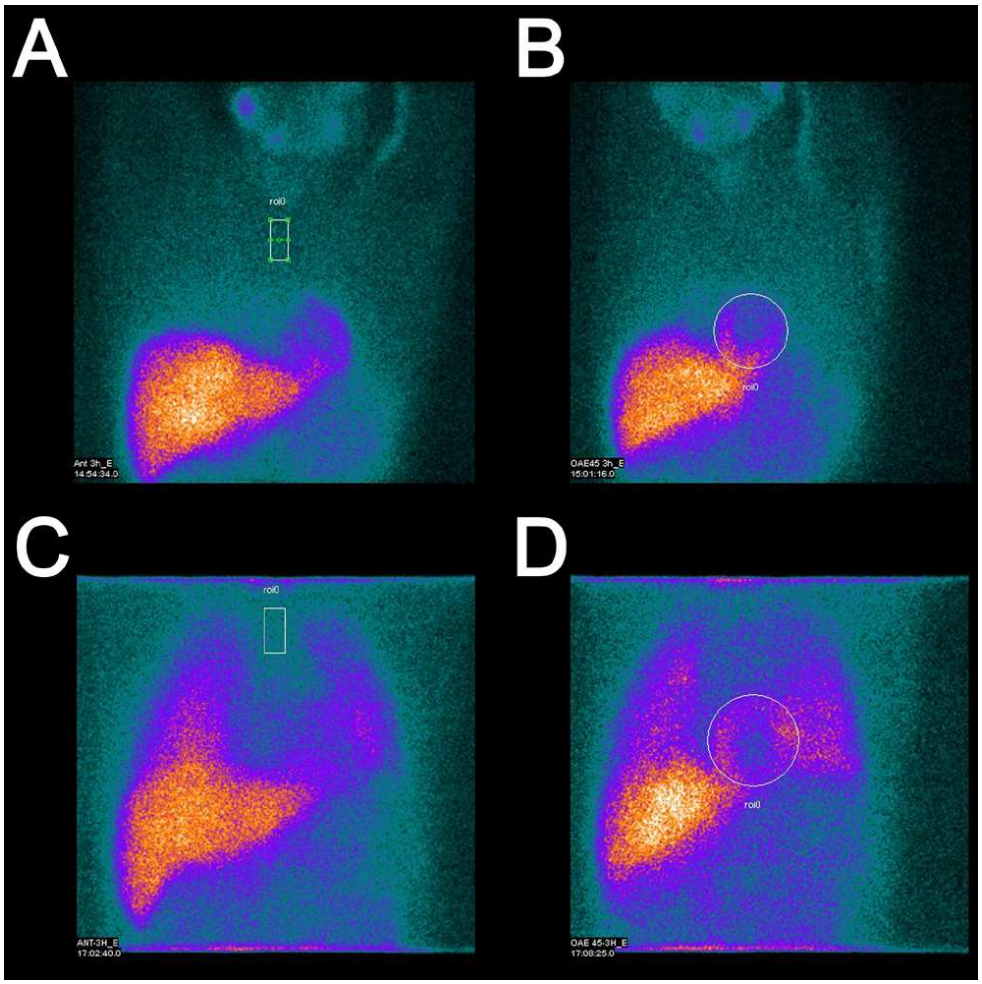


97x96mm (300 x 300 DPI)

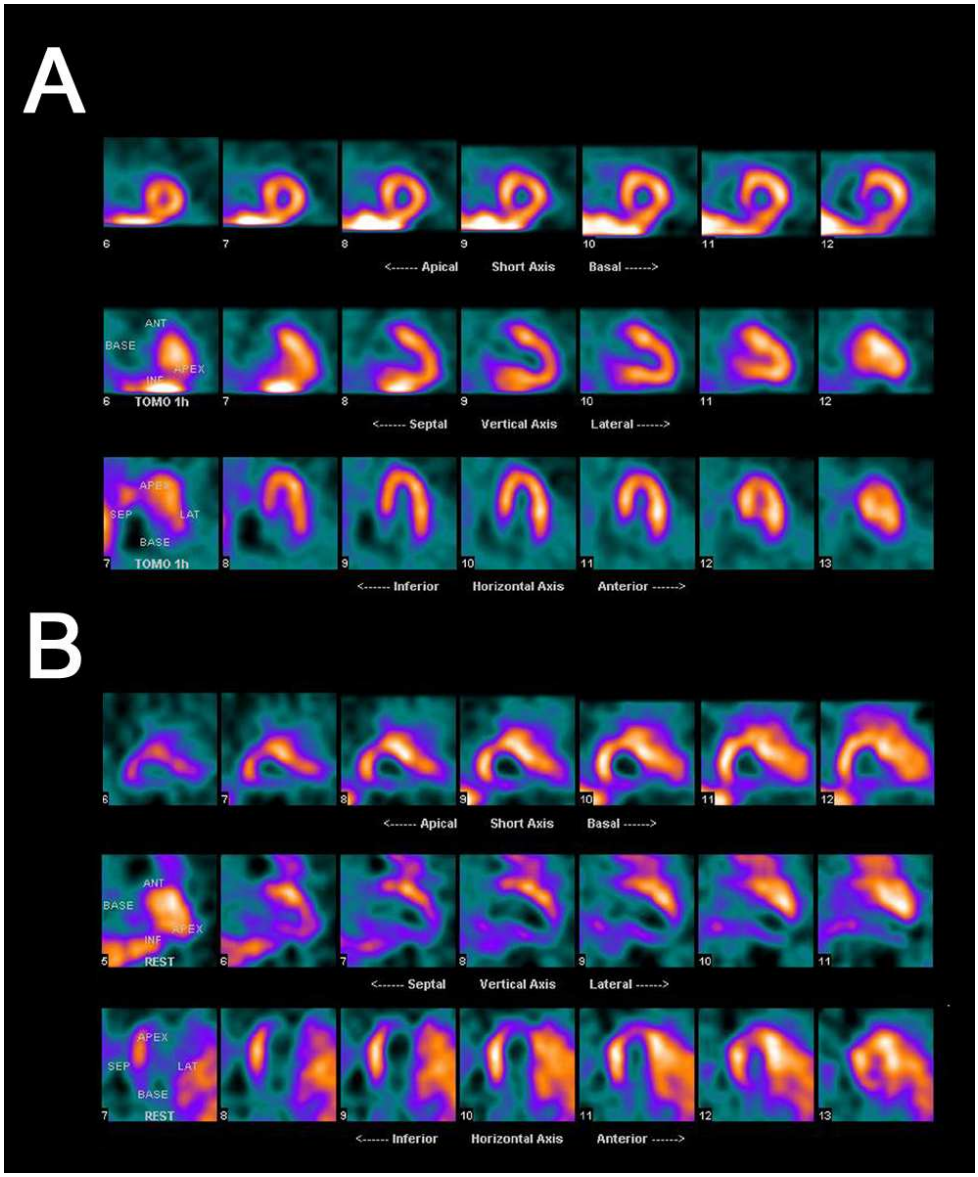


90x90mm (300 x 300 DPI)


3 h H/M ratio
2.5
2.0
1.5
normal ECG
“borderline” ECG
control



83x83mm (300 x 300 DPI)



83x101mm (300 x 300 DPI)

Apêndice I

Revisão de Estudos e Metodologias para a Associação entre Poluição Atmosférica e Doenças Respiratórias

O artigo intitulado "Artificial neural networks: a tool for evaluating the health effects of atmospheric pollution" faz uma revisão de estudos realizados e das metodologias aplicadas para a determinação dos principais fatores relacionados à poluição atmosférica que causam doenças do trato respiratório. Esses estudos utilizam dados de internações hospitalares, visitas a salas de emergência e mortalidade. O artigo foi submetido, em agosto de 2010, para publicação no "Journal of Environmental Research", encontrando-se sob a avaliação dos revisores.

Artificial neural networks: A tool for evaluating the health effects of atmospheric pollution

Emília Matos do Nascimento[a,*], José Manoel de Seixas[b], Basilio de Bragança Pereira[a,c]

[a] *Federal University of Rio de Janeiro, COPPE - Postgraduate School of Engineering, Rio de Janeiro, Brazil*
[b] *Federal University of Rio de Janeiro, Signal Processing Laboratory - COPPE/POLI, Rio de Janeiro, Brazil*
[c] *Federal University of Rio de Janeiro, School of Medicine and HUCFF - University Hospital Clementino Fraga Filho, Rio de Janeiro, Brazil*

Abstract

Air pollution is one of environmental problems with most harmful health effects. Several studies have been conducted to identify the association between weather conditions and air pollution with mortality or hospital admissions due to breathing diseases. Most researchers have applied the Generalized Additive Models and the Generalized Linear Models. Despite being often used to forecast pollutants concentration in the atmosphere, few studies apply artificial neural network models to assess the association between air pollution and its adverse health effects. This paper reviews methodologically the works performed to evaluate the health effects of atmospheric pollution, emphasizing the neural network approach.

Keywords: Air pollution; Artificial neural networks; Exposure; Greenhouse gases; Respiratory diseases

1. Introduction

The human action contributes to the accumulation of solid, liquids or gaseous residues in the environment (Bates et al., 2008). There is strong

*Corresponding author. Tel.: +55-21-25622594.
Email address: emilia@pep.ufrj.br (Emília Matos do Nascimento)

evidence of impacts of air pollution on health.

Outdoor pollutants are considered to play a role in exacerbating asthma (Atkinson and Strachan, 2004), increasing the number of hospital admissions and emergency room visits in periods with higher outdoor air pollution (Künzli et al., 2009).

The increasing number of vehicles and industries in large urban centers has exacerbated the air pollution problem. Traffic-related air pollution may increase prevalence of asthma symptoms and has been associated with reduced lung function, especially expiratory flows in schoolchildren (Rosenlund et al., 2009).

Exposure to indoor air pollution from solid fuels may be a causal agent of diseases such as acute respiratory infections and otitis media, chronic obstrutive pulmonary disease (COPD), lung cancer from coal smoke, asthma, cancer of nasopharyns and laryns, tuberculosis, perinatal conditions and low birth weight, cataract and blindness (Ezzati and Kammen, 2002; Hu and Ran, 2009). Air pollution caused by coal combustion directly affects tuberculosis incidence (Tremblay, 2007).

In developing countries, indoor air pollution may increase the risk of COPD and acute respiratory infectious in childhood, low birth weight, increase infant and perinatal mortality. Therefore it requires greatly increased efforts in the areas of research and policy-making (Bruce et al., 2000).

Recently, EPA's (Environmental Protection Agency) endangerment has pointed out the mix of atmospheric concentrations of six key greenhouse gases that threat the public health and welfare of current and future generations: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and sulfur hexafluoride (SF_6). Once emitted, they remain in the atmosphere for decades to centuries. In addition, EPA's endangerment states that the combined emissions of CO_2, CH_4, N_2O, and HFCs from new motor vehicles and motor vehicle engines contribute to threat of climate change (US Environmental Protection Agency, 2009b).

In order to identify significant associations between the air pollution and its adverse health effects, several studies have been developed, in searching for models that might fit the data and thus provide a better understanding of the air pollution effects in health.

This work aims at reviewing the studies that have been conducted to determine the main factors that cause respiratory diseases due to air pollution, considering hospitalization, visits to emergency rooms and mortality data. We focus on the methodologies that have been applied in such studies.

Particular attention, here, is given to artificial neural network models, as this computational intelligence technique (Haykin, 1999) has proved to be a successful feature extractor in a variety of pattern recognition problems from different research fields (Lamers et al., 1998; Nutman et al., 1998; Wang et al., 2008).

2. Studies linking air pollution and health effects

The present development model of technological advanced societies based on burning fossil fuels is causing the emission of greenhouse gases to Earth's atmosphere in increasing quantities, with tragic consequences. Many studies have been evaluated in order to identify these consequences, especially in health, associating each air pollutant with human disease. Table 1 summarizes the main features and adverse health effects of the of six key greenhouse gases and other atmospheric pollutants (US Environmental Protection Agency, 2010, 2009a,b,c; Griffin, 2007).

Air pollution is identified as one of the main responsible factors for the occurrence of deaths and hospital admissions for respiratory diseases. It is estimated that 42% of the cases of chronic obstructive pulmonary disease (COPD), which consists of a gradual loss of lung function, are due to environmental risk factors such as exposure to dust and chemicals, in addition to indoors air pollution (Prüss-Üstün and Corvalán, 2006). Guttorp (2000) presented a list of reseach areas in environmental statistics that need methodological developments.

Significant associations between morbidity and mortality due to respiratory diseases in urban populations and air pollution have been identified by several researchers. Table 2 presents a summary of such studies. Countless studies of time series have identified a positive association between PM concentrations, morbidity and mortality (Stieb et al., 2002; Pope et al., 1995 apud Peng et al., 2006; Bell et al., 2004).

Several researches have investigated the relationship between temperature and mortality, which is generally positive in the warm days of summer and negative for cold days of winter (Curriero et al., 2002 apud Peng et al., 2006). An approach to adjust the effects of the temperature confounding variables consists of including, in the model, nonlinear functions of the temperature of the current day and of the previous days. Peng et al. (2006) analyzed the problem of controlling the effects of confounding variables such as seasonality and trends in the long term.

Table 1: Principal atmospheric pollutants and their health effects

Pollutant	Characteristic / origin	Principal health effects
Carbon dioxide (CO_2)	It is a naturally occurring gas, and also a by-product of burning fossil fuels and biomass, as well as land-use changes and other industrial processes.	Different effects according to the concentration: above 2%, strong effect on respiratory physiology; above 7-10%, unconsciousness and death; greater than 17%, loss of controlled and purposeful activity, unconsciousness, convulsions, coma, and death occur within 1 minute of initial inhalation.
Methane (CH_4)	It is a class of greenhouse gas emitted from motor vehicles. Emissions are a function of the methane content of the motor fuel, the amount of hydrocarbons passing uncombusted through the engine, and any post-combustion control of hydrocarbon emissions.	At high enough concentrations, it is asphyxiant, capable of displacing enough oxygen to cause death by suffocation.
Nitrous oxide (N_2O)	It is produced by soil cultivation practices, especially the use of commercial and organic fertilizers, fossil fuel combustion, nitric acid production, and biomass burning.	Even at lower concentrations, it causes central nervous system, cardiovascular, hepatic, hematopoietic, and reproductive effects.
Hydrofluorocarbons (HFC)	Emissions result from the use of HFCs in cooling systems designed for passenger comfort, and auxiliary systems for refrigeration.	Cardiac sensitization, anesthesia or central nervous system related effects, malignant hyperthermia, and hepatotoxicity.
Sulfur hexafluoride (SF_6)	It is a colorless gas soluble in alcohol and ether, slightly soluble in water and used primarily in electrical transmission and distribution systems and as a dielectric in electronics.	It is asphyxiant at high concentrations.
Carbon monoxide (CO)	It is an odorless, colorless, tasteless gas, resulting from the incomplete burning of carbon-containing fuels and photochemical reactions in the atmosphere.	The effects are related to the ability to transport oxygen to the blood. It also produces heart and lung diseases, slow reflexes, fatigue and headaches.
Nitrogen dioxide (NO_2)	It is emitted from both natural (biogenic) and anthropogenic processes, such as on-road and non-road mobile sources; electricity-generating units; industrial, commercial, and residential fuels; and industrial processes.	Damages to the cell membranes in the lung tissues and constriction of the airway passages. In asthmatics, causes edema, or a filling of the intercellular spaces with fluid.
Ozone (O_3)	It is produced by natural sources as well as motor vehicle exhaust and industrial emissions, gasoline vapors, and chemical solvents.	Irritation to the eyes, precocious aging, headache, cough, chest pain, lung function losses, asthma attacks, decrease in organic resistance to infections.
Particulate matter (PM)	It is originated from a variety of anthropogenic stationary and mobile sources as well as from natural sources such as the smoke of diesel vehicles; burnings; road works; dust from public roads and industrial waste. PMs are classified according to their diameter.	Irritation of the respiratory tract, asthma, chronic bronchitis, losses of lung capacity and lung damage. In addition, it acts as carriers for toxic contaminants.
Sulfur dioxide (SO_2)	It is resulting from the burning of sulfur-based fuels, in particular, the diesel oil. In combination with water, it turns into sulfuric acid, which is the main substance in the formation of acid rain.	Irritation of the respiratory system, cough, short of breath, bronchitis, and reduction of the organic resistance to infections.

Table 2: Studies that search for associations between air pollution and health effects

Study	Period / place	Subject	Group	Methods	Significant association
Bell et al. (2004)	Several	Morbidity and mortality	Several	Review (GAM and GLIM)	PM
Dominici et al. (2003)	Several	Morbidity and mortality	Several	Review and comparison of epidemiological designs and statistical approaches	Several
Peng et al. (2006)	1987-1994 / Minneapolis, St. Paul (USA)	Mortality (all causes)	All ages	GAM and GLIM (simulated data)	PM_{10}
Singer et al. (2002)	1994 - 1997 / São Paulo, SP, Brazil	Mortality (respiratory diseases)	Children	GAM	CO
Stieb et al. (2002)	Several	Morbidity and mortality	Several	Meta-analysis (109 papers)	PM_{10}, CO, NO_2, O_3, and SO_2
Willems et al. (2007)	1997 - 2000 / Paris	Hospital admissions (infant bronchiolitis)	Children	GAM	PM_{10}

The meteorological variables and the co-pollutants can function as confounding variables in the association between exposure to PM and mortality. For example, the cold temperatures of winter, which usually result in a higher concentration of PM due to the increasing use of energy, may be considered confounding variables. Moreover, stable atmospheric conditions caused by thermal inversions are common in colder temperatures, causing higher pollutant concentration (Bell et al., 2004).

Co-pollutants such as CO, NO_2, SO_2, and O_3 are potential confounding variables. Their concentrations can be correlated with the PM due to common sources and meteorological factors. SO_2 and NO_2 are also related to the formation of the PM, contributing to patterns of similar concentration (US Environmental Protection Agency, 2003).

Techniques used to adjust the effects of confounding variables of the current and previous temperature levels included the use of moving average of previous temperatures, stratification by temperature, correction for periods of heat waves and comparison of regions with similar climatic conditions, but with different pollutant concentration levels (Bell et al., 2004; Peng and Dominici, 2008).

Bell et al. (2004) did a review of methodologies and results for time series analysis, estimating the health risks associated with exposure to PM in

the short term, noting that much of the discussion refers to epidemiological studies of air pollution in general.

They examined the critical role of those studies on policies to control air pollution and epidemiological history of PM, linking it to a bunch of adverse health effects, such as the increasing number of hospital admissions and visits to emergency rooms, exacerbation of chronic respiratory and cardiovascular diseases, pulmonary failure, and premature mortality.

Mortality data in daily time series may also be influenced by trends in population survival (including the larger or smaller access to medical care), changes in the population size, and trends in the occurrence of main diseases. These trends in long-term may coincide with recent reductions in several pollution indicators such as total suspended particles and PM_{10} (Peng et al., 2006), which are particles with aerodynamic diameter $\leq$ 10 micrometers (μm).

Regression models have been used in time series analysis to estimate the increasing risk of a health problem. For example, the mortality associated with the increase in air pollution levels in the short term. Statistical methods for time series analysis often include regression models with smooth functions of time and temperature to adjust the seasonal variations, trends in long-term, and temporal changes in temperature, that may influence the estimation of the health risk (Bell et al., 2004; Peng and Dominici, 2008).

The generalized linear models (GLIM) (McCullagh and Nelder, 1989) and the generalized additive models (GAM) (Hastie and Tibshirani, 1990) are the regression models usually chosen. The GLIM approaches include simple linear and multiple regression models, logistic regression, Poisson regression, and other models such as the log-linear models for categorized data (Nelder and Wedderburn, 1972). The GAM allow the adjustment of the models without parametric specification of the relations between independent and dependent variables. The adjustment is based on non-parametric functions, known as smoothing curves, where the type of association is defined by the data.

For instance, Willems et al. (2007) used GAM to estimate short term effects of air pollution on infant bronchiolitis hospital consultations in Paris. They found that PM_{10} was the covariate that most influenced the counts of hospital admissions.

3. Neural network models

Artificial neural networks have been widely used to predict pollutant concentrations in the atmosphere. The neural models usually used in such applications are based on single hidden layer multilayer perceptron (MLP) architectures (Haykin, 1999). Data from air pollution and weather conditions usually feed the input nodes of the network. The output node is usually related to morbidity or mortality events.

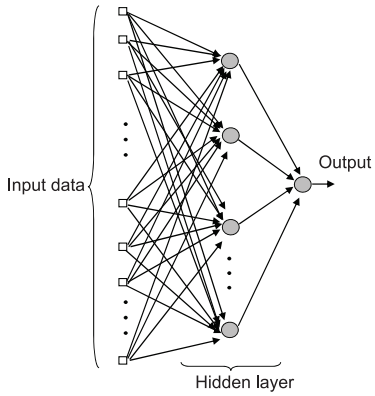


Figure 1: Architecture of the artificial neural model

Figure 1 shows the typical network topology. The gray full circles represent neurons, which gain functions may be linear or nonlinear (hyperbolic tangent). The arrows connecting each pair of neurons correspond to the weighting vectors, which are adapted through some optimizaton algorithm.

Most applications apply supervised learning, in which the target values for the output node are known a priori. The backpropagation algorithm (Haykin, 1999) is very popular, using the error values measured at the output node to adapt both output and hidden layer weights.

During the implementation, the need for data pre-processing may be evaluated (Nelson et al., 1999). In time series applications, data pre-processing consists of the removal of trend, seasonality and moving averages. The trend may be estimated by the least squares estimator. The seasonality may be modelled by Fourier Analysis (Cooley and Tukey, 1965), which may be used to identify the cycles to be removed. A relevance study may also be used as a selection criterion for the explanatory variables (Seixas et al., 1996).

Gardner and Dorling (1998) did a literature review and found many applications of the multilayer perceptron (MLP) neural model for forecasting,

Table 3: Studies related to air pollution and health effects applying artificial neural networks

Study	Period / local monitoring	Subject	Group	Methods	Significant association
Bibi et al. (2002)	1992-1995 / Barzilai Medical Center (Ashkelon, Israel)	Prediction of emergency department visits for respiratory symptoms	Patients with symptoms of asthma, COPD, and acute and chronic bronchitis	Back-propagation training algorithm and genetic algorithm	SO_2, NO_x, and weather conditions
Lamers et al. (1998)	Winter of 1991 and 1992 / Utrecht, Netherlands	Predicting pulmonary response to air pollution	Schoolchildren aged 7-12	Feedforward network models	SO_2, NO_2, and black smoke
Moseholm et al. (1993)	1987-1988 / Odense and Aalborg, Denmark	Response in subjects with asthma to gaseous air pollution levels, weather, and medicine intake	Outpatients aged 18-60 who had been diagnosed as having asthma the previous year	Back-propagation training algorithm and principal component analysis	SO_2 and NO_2
Nascimento et al. (2009)	1997 - 2000 / Paris	Hospital admissions due to infant bronchiolitis	Children	Feedforward network models	PM_{10}
Nutman et al. (1998)	September 1993 - November 1993 / Tel-Aviv, Israel	Pollution levels and bronchial asthma-associated emergency room visits	Adults aged 15-75	Back-propagation training algorithm	NO_x
Wang et al. (2008)	2000-2005 / Beijing, China	Mortality due to respiratory diseases	All ages	Granger causality and Back-propagation training algorithm	NO_x

function approximation, and pattern classification in problems where weather and air pollution were involved.

However, unlike the popularity of GAM and GLIM approaches, only few studies apply the neural models in searching for the association between air pollution and its harmful effects to health. Table 3 presents some of such studies.

Nutman et al. (1998) used a neural model trained by backpropagation algorithm (Haykin, 1999) to examine the relationship between air pollution and visits by adults to emergency rooms of a hospital in Tel-Aviv, Israel, due to asthma. According to the authors, the results clearly showed that the levels of air pollution were mainly due to the NO_x concentration, which consists of one molecule of nitrogen and varying numbers of oxygen molecules

(US Environmental Protection Agency, 2009a).

Bibi et al. (2002) developed a neural model based on backpropagation and Genetic Algorithm (Goldberg, 1989) to predict the number of visits of patients with symptoms of asthma, bronchitis, and COPD, to the emergency department of the Barzilai Medical Center in Ashkelon, Israel. Genetic algorithms (Holland, 1992) are algorithms of optimization and search based on the mechanisms of natural selection and genetics. In that application they were used to determine the number of neurons in the hidden layer (Miller et al., 1989 apud Bibi et al., 2002).

A study by Wang et al. (2008) applied Granger causality (Granger, 1969) and artificial neural networks to investigate the relationship between air pollution and mortality due to respiratory diseases in Beijing, China. They concluded that NO_x was the most important pollutant to control.

Lamers et al. (1998) analyzed a dataset from schoolchildren aged 7-12 in Utrecht, Netherlands, to predict pulmonary response to air pollution. Firstly, the authors modelled the data with standard feedforward network. Afterwards, they used combined neural network models for reducing correlation between covariates to improve prediction quality.

Moseholm et al. (1993) applied artificial neural networks and Principal Component Analysis (Jolliffe, 2002) to study the response in asthmatics to gaseous air-pollution levels, weather, and medicine intake in Odense and Aalborg, Denmark.

Nascimento et al. (2009) have recently applied artificial neural networks to analyze the impact of air pollution and weather conditions on the hospital admissions due to infant bronchiolitis, making use of the same dataset modeled by Willems et al. (2007). The neural models were evaluated in terms of data description, when all the information was used in its development, or were designed aiming at generalization. During the implementation, the need for data pre-processing was evaluated. The results agreed with the previous GAM-based approach (see Table 2), pointing out that PM_{10} was the explanatory variable with largest influence in the counts of hospital admissions.

4. Neural networks and statistics

Many neural models are similar to conventional statistical methods, such as the GLIM, polynomial regression, non-parametric regression, and discriminant analysis, principal component analysis, and cluster analysis (Detienne

et al., 2003).

A multilayer feedforward neural network trained by backpropagation algorithm is a kind of non-parametric regression model. The evaluation of the regression coefficients is done using a data set composed of independent variable associated with known values of the dependent variables. In a similar way, in supervised learning, this problem is equivalent to a feedforward neural network of a single layer, where the independent variables are the inputs and the dependent variable is the output (target); the activation function is the identity function; and the coefficients β_i of the regression correspond to the neural network weights. These weights are evaluated using the backpropagation algorithm (Warner and Misra, 1996).

The neural model differs from the regression model by using an iterative process. In multiple linear regression models, it is assumed that the output is linked to a linear combination of the independent variables. If the given model is not appropriated, the prediction will be incorrect even if the error adjustment is small. Alternatively, one can relax the assumption of any functional relationship and let the data define its functional shape, that is let the data speak for themselves. This is the power of neural networks (Warner and Misra, 1996; Detienne et al., 2003).

Regression models have some disadvantages: the relationship between dependent and independent variables must be identified in advance; the prior assumptions of the regression parameters; the regression models are not adaptive; the need of a priori knowledge of the data distribution; and non-recognition of multicolinearity, which is the high correlation degree between two or more variables (Detienne et al., 2003).

In turn, neural modeling has also some disadvantages: neural networks do not reveal immediately the functional relationship between variables; and it is not easy to choose parameters as the number of neurons in the hidden layer, the learning rate (η), initial weights, objective function, and the stop training decision. These parameters are usually determined empirically (Warner and Misra, 1996).

5. Conclusion

The concern about the air pollution effects has been the subject of discussion between leaders of different countries, leading them to sign agreements to control and reduce the greenhouse gas emissions (UNFCCC, 1992, 1998).

The leaders of the Group of 8 countries agreed that the greenhouse gas emissions should be reduced "to limit the average increase in global temperature to 2 degrees Celsius above pre-industrial levels". However, the developing countries do not agree to limit their growth to fix damage caused by industrial countries (Baker, 2009). It is imperative that health authorities take action related to policies to be adopted in each area where the air pollutant effects are harmful.

This paper provided a methodological review of the studies involving air pollution and health effects. Although neural models have not been generally used in the search for association between air pollution and its health effects, their growing use shows that they represent a powerful tool in the search for such association. Despite the technique, large number of models already developed clearly shows that public and private health systems and life itself suffer a tremendous impact from atmospheric pollution.

Acknowledgements

We would like to thank FAPERJ, CNPq and CAPES (Brazil) for their support to this work.

References

Atkinson, R.W., Strachan, D.P., 2004. Role of outdoor aeroallergens in asthma exacerbations: Epidemiological evidence. Thorax 59, 277–278.

Baker, P., 2009. Poorer nations reject a target on emission cut. New York Times. Retrieved on July 18, 2009, from `http://www.nytimes.com/2009/07/09/world/europe/09prexy.html?pagewanted=1&2334&2359&_r=2&sq&st=Search&2359;2degree&2359;&scp=3`.

Bates, B.C., Kundzewicz, Z.W., Wu, S., Palutikof, J.P., 2008. Paper of the Intergovernmental Panel on Climate Change. Technical Report. IPCC. Geneva. Retrieved on March 19, 2009, from `http://www.ipcc.ch/pdf/technical-papers/climate-change-water-en.pdf`.

Bell, M.L., Samet, J.M., Dominici, F., 2004. Times-series studies of particulate matter. Annu Rev Public Health 25, 247–280.

Bibi, H., Nutman, A., Shoseyov, D., Shalom, M., Peled, R., Kivity, S., Nutman, J., 2002. Prediction of emergency department visits for respiratory symptoms using an artificial neural network. Chest 122, 1627–1632.

Bruce, N., Perez-Padilla, R., Albalak, R., 2000. Indoor air pollution in developing countries: A major environmental and public health challenge. Bull World Health Organ 78, 1078–1092.

Cooley, J., Tukey, J., 1965. An algorithm for the machine calculation of complex fourier series. Mathematics of Computation 19, 297–301.

Curriero, F.C., Heiner, K.S., Samet, J.M., Zeger, S.L., Strug, L., Patz, J.A., 2002. Temperature and mortality in 11 cities of the eastern United States. Am J Epidemiol 155, 80–87.

Detienne, K.B., Detienne, D.H., Joshi, S.A., 2003. Neural networks as statistical tools for business researchers. Organizational Research Methods 6, 236–265.

Dominici, F., Sheppard, L., Clyde, M., 2003. Health effects of air pollution: A statistical review. Int Stat Rev 71, 243–276.

Ezzati, M., Kammen, D.M., 2002. The health impacts of exposure to indoor air pollution from solid fuels in developing countries: Knowledge, gaps, and data needs. Environ Health Perspect 110, 1057–1068.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. Atmos Environ 32, 2627–2636.

Goldberg, D.E., 1989. Genetic algorithms in search, optimization and machine learning. Addisson Weslley.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Griffin, R.D., 2007. Principles of air quality management. CRC Press, Boca Raton.

Guttorp, P., 2000. Environmental statistics. J Am Stat Assoc 95, 289–292.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models. Chapman and Hall, London.

Haykin, S., 1999. Neural networks: A comprehensive foundation. Prentice Hall. 2 edition.

Holland, J.H., 1992. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence. MIT Press.

Hu, G., Ran, P., 2009. Indoor air pollution as a lung health hazard: Focus on populous countries. Curr Opin Pulm Med 15, 158–164.

Jolliffe, I.T., 2002. Principal component analysis. Springer, New York.

Künzli, N., Bridevaux, P.O., Liu, L.J.S., Garcia-Esteban, R., Schindler, C., Gerbase, M.W., Sunyer, J., Keidel, D., Rochat, T., 2009. Traffic-related air pollution correlates with adult-onset asthma among never-smokers. Thorax 64, 664–670.

Lamers, M.H., Kok, J., Lebret, E., 1998. Combined neural network models for epidemiological data: Modelling heterogeneity and reduction of input correlation, in: Smith, G., Steele, N.C., Albrecht, R. (Eds.), Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms, Springer-Verlag. pp. 147–151.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models. Chapman and Hall, London.

Miller, G.F., Todd, P.M., Hegde, S., 1989. Designing neural networks using genetic algorithms, in: Schaffer, J.D. (Ed.), Proceedings of the third international conference on Genetic Algorithms, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 379–384.

Moseholm, L., Taudorf, E., Frøsig, A., 1993. Pulmonary function changes in asthmatics associated with low-level SO_2 and NO_2 air pollution, weather, and medicine intake: An 8-month prospective study analyzed by neural networks. Allergy 48, 334–344.

Nascimento, E.M., Pereira, B.B., Seixas, J.M., 2009. Artificial neural networks: an application in the study of air pollution and its adverse health effects. Biometric Brazilian Journal 27, 37–50. In Portuguese.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J R Stat Soc Ser A 135, 370–384.

Nelson, M., Hill, T., Remus, W., O'Connor, M., 1999. Times series forecasting using neural networks: Should the data be deseasonalized first? J Forecast 18, 359–367.

Nutman, A., Solomon, Y., Mendel, S., Nutman, J., Hines, E., Topilsky, M., Kivity, S., 1998. The use of a neural network for studying the relationship between air pollution and asthma-related emergency room visits. Respir Med 92, 1199–1202.

Peng, R.D., Dominici, F., 2008. Statistical methods for environmental epidemiology with R: A case study in air pollution and health. Springer, New York.

Peng, R.D., Dominici, F., Louis, T.A., 2006. Model choice in times-series studies of air pollution and mortality. J R Stat Soc Ser A 169, 179–203.

Pope, C.A., Dockery, D., Schwartz, J., 1995. Review of epidemiological evidence of health effects of particulate air pollution. Inhal Toxicol 47, 1–18.

Prüss-Üstün, A., Corvalán, C., 2006. Preventing disease through healthy environments: Towards an estimate of the environmental burden of disease. World Health Organization, Geneva. Retrieved on July 14, 2009, from `http://www.who.int/quantifying_ehimpacts/publications/preventingdisease/en/index.html`.

Rosenlund, M., Forastiere, F., Porta, D., De Sario, M., Badaloni, C., Perucci, C., 2009. Traffic-related air pollution in relation to respiratory symptoms, allergic sensitisation, and lung function in schoolchildren. Thorax 64, 573–580.

Seixas, J.M., Calôba, L.P., Delpino, I., 1996. Relevance criteria for variance selection in classifier designs, in: International Conference on Engineering Applications of Neural Networks, London. pp. 451–454.

Singer, J.M., Andre, C.D.S., Lima, L.P., Conceição, G.M.S., 2002. Atmospheric pollution and mortality in São Paulo, in: Dodge, Y. (Ed.), Statistical data analysis based on the L1-norm and related methods, Birkhäuser Verlag, Basel, Switzerland. pp. 439–450.

Stieb, D.M., Judek, S., Burnett, R.T., 2002. Meta-analysis of time-series studies of air pollution and mortality: Effects of gases and particles and the influence of cause of death, age, and season. J Air Waste Manage Assoc 52, 470–484.

Tremblay, G.A., 2007. Historical statistics support a hypothesis linking tuberculosis and air pollution caused by coal. Int J Tuberc Lung Dis 11, 722–732.

UNFCCC, 1992. United Nations framework convention on climate change. Retrieved on July 18, 2009, from http://unfccc.int/essential_background/convention/background/items/2853.php.

UNFCCC, 1998. The Kyoto protocol to the framework convention on climate change. Retrieved on July 18, 2009, from http://unfccc.int/essential_background/kyoto_protocol/background/items/1351.php.

US Environmental Protection Agency, 2003. Air quality criteria for particulate matter (fourth external review draft, jun 2003). (EPA/600/P-99/002aD and bD). Retrieved on August 5, 2010, from http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=58003.

US Environmental Protection Agency, 2009a. Glossary of climate change terms. Retrieved on August 22, 2009, from http://www.epa.gov/climatechange/glossary.html.

US Environmental Protection Agency, 2009b. Proposed endangerment and cause or contribute findings for greenhouse gases under section 202(a) of the clean air act. (EPA-HQ-OAR-2009-0171). Retrieved on August 18, 2009, from http://www.regulations.gov/search/Regs/contentStreamer?objectId=090000648096894b&disposition=attachment&contentType=pdf.

US Environmental Protection Agency, 2009c. Technical support document for endangerment and cause or contribute finding for greenhouse gases under section 202(a) of the clean air act. Retrieved on August 18, 2009, from http://epa.gov/climatechange/endangerment/downloads/TSD_Endangerment.pdf.

US Environmental Protection Agency, 2010. Air quality: Epa's integrated science assessments. Retrieved on August 5, 2010, from `http://www.epa.gov/ncea/isa/`.

Wang, Q., Liu, Y., Pan, X., 2008. Atmosphere pollutants and mortality rate of respiratory diseases in beijing. Sci Total Environ 391, 143–148.

Warner, B., Misra, M., 1996. Understanding neural networks as statistical tools. Am Stat 50, 284–293.

Willems, S., Segala, C., Maidenberg, M., Mesbah, M., 2007. Longitudinal analysis of short-term bronchiolitis air pollution association using semi-parametric models, in: Auget, J.L., Balakrishnan, N., Mesbah, M., Molenberghs, G. (Eds.), Advances in Statistical Methods for the Health Sciences: Applications to cancer and AIDS studies, genome, sequence analysis, and survival analysis, Birkhäuser, Boston. pp. 467–487.