UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Engenharia Química

TIAGO DIAS MARTINS

PREDIÇÃO DA RECORRÊNCIA DE TROMBOEMBOLISMO VENOSO VIA REDES NEURAIS ARTIFICIAIS

Campinas - São Paulo
2018

TIAGO DIAS MARTINS

PREDIÇÃO DA RECORRÊNCIA DE TROMBOEMBOLISMO VENOSO VIA REDES NEURAIS ARTIFICIAIS

Tese apresentada à Faculdade de Engenharia Química da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Engenharia Química

Orientador: Prof. Dr. Rubens Maciel Filho
Coorientadora: Profa. Dra. Joyce Maria Annichino-Bizzacchi

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO TIAGO DIAS MARTINS, ORIENTADA PELO PROF. DR. RUBENS MACIEL FILHO E CO-ORIENTADA PELA PROF. DRA. JOYCE MARIA ANNICHINO-BIZZACCHI

Campinas - São Paulo
2018

Agência(s) de fomento e nº(s) de processo(s): CAPES, 33003017034P8
ORCID: https://orcid.org/0000-0002-3452-703

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

M366p

Martins, Tiago Dias, 1986-
Predição da recorrência de tromboembolismo venoso via redes neurais artificiais / Tiago Dias Martins. – Campinas, SP : [s.n.], 2018.

Orientador: Rubens Maciel Filho.
Coorientador: Joyce Maria Annichino-Bizzacchi.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Química.

1. Tromboembolia. 2. Inteligência artificial. 3. Análise estatística multivariada. I. Maciel Filho, Rubens, 1958-. II. Annicchino-Bizzachi, Joyce Maria, 1957-. III. Universidade Estadual de Campinas. Faculdade de Engenharia Química. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Prediction of recurrent thromboembolism using artificial neural networks
Palavras-chave em inglês:
Thromboembolism
Artificial intelligence
Multivariate statistical analysis
Área de concentração: Engenharia Química
Titulação: Doutor em Engenharia Química
Banca examinadora:
Rubens Maciel Filho [Orientador]
Marina Pereira Colella
Igor Tadeu Lazzarotto Bresolin
Roberto Nasser Junior
Edvaldo Rodrigo de Morais
Data de defesa: 10-04-2018
Programa de Pós-Graduação: Engenharia Química

Tese de Doutorado defendida por Tiago Dias Martins e aprovada em 10 de abril de 2018 pela banca examinadora constituída pelos doutores.

Prof. Dr. Rubens Maciel Filho (orientador)

Prof. Dr. Edvaldo Rodrigo de Morais

Prof. Dr. Igor Tadeu Lazzarotto Bresolin

Profa. Dra. Marina Pereira Colella

Prof. Dr. Roberto Nasser Junior

A ata da defesa com as respectivas assinaturas dos membros da banca examinadora encontra-se no processo de vida acadêmica do aluno.

“I think our Universe isn't just described by Math. I think it IS Math. I think our entire Universe is a giant mathematical structure that we are part of.”

Prof. Dr. Max Tegmark – MIT

AGRADECIMENTOS

Agradeço, primeiramente, a meus pais. Sem o esforço deles eu não chegaria onde estou.

Agradeço ao meu companheiro Rafael, por ter me incentivado a continuar esta longa jornada.

Agradeço aos colegas do Departamento de Engenharia Química, da Universidade Federal de São Paulo, por todo o incentivo e o apoio para que eu pudesse finalizar esta Tese de Doutorado.

Agradeço a todos os meus mentores, orientadores, que passaram ao longo de toda a minha trajetória:

> Ao Prof Dr. Edson Antonio da Silva, que me ensinou, me incentivou a usar as redes neurais artificiais e que despertou em mim o gosto pela pesquisa científica.
>
> Ao Prof Dr. Charlles Rubber de Almeida Abreu, que orientou meu trabalho de mestrado.
>
> À Profa. Dra. Marisa Masumi Beppu, que me acolheu em seu laboratório, me apresentou o lado experimentalista da pesquisa, contribuiu significativa para a minha formação como pesquisador, mas especialmente para a minha formação como Professor.
>
> Ao Prof. Dr. Rubens Maciel Filho, que me acolheu como seu orientado nos últimos anos e que me apresentou a essa linha de pesquisa inovadora e de extrema importância para a sociedade. Obrigado pelo seu tempo e sua grande contribuição para a minha carreira.
>
> À Profa. Joyce Maria Annichino-Bizzacchi, que aceitou o desafio da interdisciplinariedade entre duas áreas tão distintas. Obrigado pelas reuniões e pelas frutíferas discussões que tivemos.

Se fui aprovado em um concurso para Professor em uma Universidade Federal antes mesmo de concluir o Doutorado, devo muito às oportunidades que vocês me deram para crescer.

Agradeço, em especial, à colega Anna Calazans. Sem ela, este estudo não teria acontecido. Obrigado pela disposição e por todo seu empenho.

Agradeço à FAPESP e à Capes, pelo auxílio financeiro ao desenvolvimento desse trabalho.

RESUMO

A recorrência da trombose venosa pode acometer até 30 % dos pacientes após um primeiro episódio, em 5 anos. Após um primeiro episódio trombótico, o tratamento padrão é a administração de um medicamento anticoagulante por 3 a 6 meses. Porém, após esse período, a probabilidade de recorrência, em um percentual de pacientes, ainda é incerta. Diversos *scores* foram desenvolvidos para o cálculo dessa probabilidade na última década, mas todos possuem diversas limitações. Dentro desse contexto, se destacam as Redes Neurais Artificiais e sua versatilidade no aprendizado e generalização. Assim, o objetivo principal deste trabalho foi obter modelos neurais para predizer quais pacientes poderão ter recorrência de trombose com base somente em dados clínicos e laboratoriais. Primeiramente, foram coletados dados de 39 fatores para 235 pacientes que apresentaram uma primeira trombose nos membros inferiores ou no sistema nervoso central. Então, foram identificados os principais fatores para trombose recorrente, empregando-se a Análise de Componentes Principais. Em seguida, diversos modelos neurais foram ajustados considerando-se diferentes conjuntos de variáveis de entrada: i) os 39 fatores, e ii) os fatores principais determinados pela Análise de Componentes Principais. Também foram propostos dois modelos alternativos cujo conjunto de entrada foi composto da resposta dada pela Análise de Componentes Principais e aspectos práticos. Os resultados mostraram que apenas variáveis do hemograma, bem como dados da primeira trombose são suficientes para se determinar se um paciente apresentará recorrência. Além disso, foi possível verificar que as redes neurais artificiais são capazes de predizer o fenômeno com precisão, atingindo coeficientes de correlação igual a 0,99999. Este trabalho mostrou que a associação de técnicas estatísticas multivariadas e inteligência artificial pode ser uma alternativa eficaz para auxiliar na predição de retrombose e das mais diversas enfermidades.

Palavras-Chave: tromboembolia; inteligência artificial; análise estatística multivariada.

ABSTRACT

Recurrent thrombosis is a disease that occurs in about 30 % of venous thromboembolism cases. When a patient presents a first thrombotic event, the main treatment is anticoagulation therapy along 3 months. After this period, the probability of recurrence is uncertain. Several statistical methods to calculate this probability were developed during the last decade, but all of them present several limitations. In this context, Artificial Neural Networks gain importance especially due its versatility and generalization capabilities. Thus, the main objective of this work was to obtain neural models to predict which patients will present recurrent thrombosis considering only clinical and laboratorial data. First, it was collected information about 39 clinical factors of 235 patients that presented a first thrombotic event in inferior members or central nervous system. Then, it was determined the recurrent thrombosis main factors using Principal Component Analysis. Several artificial neural networks structures were trained considering different input variables: i) the 39 factors, and ii) the main factors indicated by principal component analysis. Also, two alternative models were proposed considering a combination of the Principal Component Analysis and practical aspects. The results showed that only hemogram variables, as well as characteristics of the first thrombosis are sufficient to determine if a patient will present recurrent thrombosis. Besides, the artificial neural networks are capable to predict the phenomenon accurately, with correlation coefficients of 0,99999. This work showed that the association of statistical multivariate techniques and artificial intelligence models can be an efficient alternative to help clinicians predict recurrent thrombosis and also different diseases.

Keywords: thromboembolism; artificial intelligence; multivariate statistical analysis.

NOMENCLATURA

Siglas em Português

AT	antitrombina
ACP	Análise de Componentes Principais
CP	componentes principais
EP	emboliar pulmonar
FT	fator tecidual
HDL	lipoproteína de alta densidade
IMC	índice de massa corporal
LDL	lipoproteína de baixa densidade
PC	proteína C
PS	proteína S
RNA	Redes Neurais Artificiais
SAF	síndrome do anticorpo fosfolipídeo
TEV	tromboembolismo venoso
TFPI	inibidor da via do fator tecidual
TVC	trombose venosa cerebral
TR	trombose recorrente
TVP	trombose venosa profunda

Siglas em inglês

ANN	Artificial Neural Network
BMI	body mass index
HB	hemoglobin

HCT	hematrocrit
HDL	high-density lipoprotein
LDL	low-density lipoprotein
MVP	mean platelet volume
PC	principal components
PCA	Principal Component Analysis
PLT	platelet
RBC	red blood cell
RDW	red blood cell width distribution
RVTE	recurrent venous thromboembolism
VTE	venous thromboembolism
WBC	white blood cell

Letras Latinas

a	matriz autovetor
b_j	bias do neurônio j
CP	matriz das componentes principais
E	erro calculado entre a resposta da rede neural artificial e o valor real
F_{OBJ}	função objetivo
i	índice auxiliar; iteração ao longo do treinamento da rede neural artificial
I_i	entrada i do neurônio artificial
I	matriz identidade
n	número de variáveis de saída da rede neural artificial
O_j	saída do neurônio j
s	matriz variância-covariância

t	parâmetro arbitrário da rede neural artificial
w_{ij}	peso da entrada i no neurônio artificial j
W	número de neurônios na camada intermediária da rede neural artificial
x_i	entrada i da rede neural artificial
x_i	saída i da rede neural artificial
Z	número de variáveis de entrada de um neurônio artificial; número de variáveis de entrada da rede neural artificial; variável padronizada

Letras Gregas

α_j	coeficiente de ativação do neurônio j
λ	autovalor
ρ	correlação entre as variáveis originais e as componentes principais
σ	variância
μ	taxa de aprendizado da rede neural artificial; valor médio

SUMÁRIO

1 INTRODUÇÃO

1.1 Enunciado do Problema

O tromboembolismo venoso (TEV) compreende a trombose venosa profunda (TVP) e a embolia pulmonar (EP), decorrentes da presença de um trombo dentro das veias ou artérias pulmonares, respectivamente. Apesar do tratamento da trombose venosa, pode haver o desprendimento de um embolo causando a embolia pulmonar, com consequências fatais. Além disso, pode haver o aumento do trombo primário ou a formação de trombos em outros sítios venosos. Como exemplo, pode-se citar a trombose do sistema nervoso central, ou trombose venosa cerebral (TVC), que atinge as veias do cérebro.

Uma vez diagnosticado com a doença, o paciente passa por um tratamento mínimo de 3 meses com a administração de um anticoagulante oral. Após completar esse período, a continuidade do tratamento é avaliada, para prevenção de recorrência da trombose. Uma das complicações do TEV é a trombose recorrente (TR), que atinge 20 a 30 % dos indivíduos nos primeiros 5 anos após o primeiro episódio trombótico. Por esse motivo, diversos estudos se concentram em identificar quais pacientes são propensos a TR, através da determinação dos fatores que predispõem à recorrência. Porém, por ser uma doença multifatorial essa não é uma tarefa fácil. Além disso, os fatores, e a interação entre eles, podem ser diferentes dependendo da população que se considera.

Assim, a busca por novos métodos que identifiquem quais são os principais fatores, bem como maneiras de se classificar os pacientes é extremamente importante e pode trazer contribuições significativas no tratamento dos pacientes, tendo impacto direto na duração do tratamento e como consequência na sua qualidade de vida. Na última década, três métodos de *scores* foram propostos visando calcular o risco de TR a partir de dados clínicos dos pacientes. Porém, eles possuem uma série de limitações, tais como: incluir somente pacientes com TEV espontâneo, não diferenciar pacientes que se encontram no grupo de baixo risco de TR, ou não terem tido uma validação externa.

Por outro lado, as técnicas estatísticas multivariadas (como a Análise de Componentes Principais) e as técnicas de inteligência artificial (como as Redes Neurais Artificiais) têm ganhado grande destaque na literatura médica. A primeira é capaz de detectar

padrões e fatores importantes de um fenômeno analisando a variância dos fatores que o influenciam. Já, a segunda, é capaz de aprender com exemplos e modelar matematicamente diversos fenômenos a partir de tais fatores, mesmo considerando ruídos nos dados e não-lineraridade entre as variáveis.

1.2 Proposta do Trabalho

Considerando o exposto no item anterior, este trabalho surge como uma alternativa aos modelos de *scores* propostos na última década. Para isso, parte-se do princípio que a TR é uma função matemática dos seus fatores com duas respostas possíveis: "sim" ou "não", que neste trabalho serão codificadas na forma de "1" ou "-1", respectivamente. Matematicamente, a função pode ser escrita da seguinte forma:

$$TR = f\left(\text{fatores que a influenciam}\right) = \begin{cases} +1; & \text{se sim} \\ -1; & \text{se não} \end{cases} \quad (1)$$

Assim, o principal objetivo deste trabalho foi explorar, desenvolver procedimentos e, aplicar técnicas estatísticas multivariadas e redes neurais artificiais, visando obter uma nova ferramenta matemática que poderá auxiliar os médicos na tomada de decisões com respeito à continuidade do tratamento anticoagulante. A associação dessas técnicas apresenta vantagens com relação aos procedimentos anteriores pela sua abrangência, capacidade de tratar com dados ruidosos, não-linearidade e perspectiva de contínuo aprendizado com a alimentação de novos dados clínicos e laboratoriais dos pacientes. Dentro desse contexto, espera-se: i) determinar os principais fatores preditivos da recorrência da TVP; ii) obter novos modelos matemáticos, empregando redes neurais artificiais para predizer a TR.

1.3 Organização do Texto

O texto desta Tese não foi escrito de forma canônica, sendo que parte dele é apresentado na forma de dois manuscritos, que contém os resultados e discussão do estudo. Devido ao fato de serem textos resumidos, os Capítulos iniciais são dedicados à

contextualização teórica dos assuntos pertinentes. Para melhor compreensão do leitor, o texto da Tese será organizado da seguinte forma:

No Capítulo 2 serão apresentados os aspectos teóricos a respeito da coagulação do sangue, TEV, sua recorrência, assim como os fatores correlacionados e os modelos matemáticos propostos para o cálculo do risco de recorrência.

Os Capítulos 3 e 4 são dedicados aos aspectos teóricos das ferramentas utilizadas neste trabalho, que comporão o procedimento desenvolvido e aplicado. No Capítulo 3, os aspectos teóricos sobre as RNAs serão apresentados: a ideia que as originaram, estrutura e funcionamento. Aspectos sobre sua utilização, treinamento e pós-treinamento também serão abordados. No Capítulo 4, serão apresentados os aspectos teóricos sobre a Análise de Componentes Principais e sua implementação. Em todos os Capítulos também serão apresentados estudos médicos que já utilizaram as técnicas supracitadas. A presença destes Capítulos na Tese se faz necessária por ser um trabalho multidisciplinar, tendo como intenção municiar o leitor dos principais conceitos utilizados no desenvolvimento do procedimento proposto e aplicado neste trabalho, sem que seja necessária a busca em literatura especializada nos assuntos.

No Capítulo 5 será apresentada uma breve descrição do procedimento empregado, incluindo a obtenção e tratamento dos dados utilizados, bem como a construção dos modelos neurais. No Capítulo 6 serão apresentados os objetivos e os principais resultados apresentados nos Capítulos 7, 8 e 9, respectivamente. Por fim, no Capítulo 10 são apresentadas as conclusões e sugestões para trabalhos futuros que surgiram como desdobramentos desta pesquisa.

1.4 Principais Contribuições deste Trabalho

Este trabalho possui várias contribuições para o estudo da TR. A primeira delas é a aplicação de técnicas estatísticas e de inteligência artificial para esse fim. Tais técnicas têm se mostrado uma ferramenta poderosa na Engenharia Química, Economia, Administração e outras áreas. Porém sua aplicação na Medicina ainda é bastante incipiente. Soma-se a isso, o fato da simbiose dessas duas técnicas raramente ser abordada na literatura. Usualmente elas são aplicadas individualmente na resolução dos mais diversos problemas.

Também pode se citar a interação entre diferentes áreas do conhecimento: Engenharia Química e Medicina. A interdisciplinariedade, ou seja, resolver problemas de uma determinada área do conhecimento aliando-se ferramentas de outra área é uma ação cada vez mais solicitada no século XXI. Os aspectos médicos raramente são abordados de uma maneira matemática, muitas vezes devido à sua natureza complexa, de difícil mensuração e muito variável. Por outro lado, o conhecimento e as ferramentas comumente empregados na Engenharia são mais utilizados para processos industriais, ou fenômenos cujo comportamento pode ser descrito com equações de balanço de massa e energia. Desse modo, outra contribuição deste trabalho foi mostrar que, ao se aliar diferentes ferramentas matemáticas, se pode obter uma descrição simples de fenômenos complexos, tal como a TR.

Por fim, este trabalho também tem como grande contribuição o desenvolvimento de uma equação matemática que tem grande utilidade pública, especialmente a nível do Brasil. A tomada de decisão sobre a duração do tratamento do primeiro evento trombótico é difícil e envolve a análise de muitas variáveis, além dos custos para a determinação de níveis de diversos fatores importantes, tais como: dosagem de dímero-D, pesquisa de trombofilias hereditárias e adquiridas. Os modelos obtidos neste trabalho podem ter um grande impacto na redução dos custos de realização desses exames, especialmente em centros com orçamento limitado, bem como facilitar e simplificar a tomada de decisão dos médicos.

Os resultados obtidos pela técnica de Análise de Componentes Principais mostraram que a TR pode ser predita a partir de parâmetros de fácil obtenção, tais como parâmetros sanguíneos e características do indivíduo e da primeira trombose. A técnica proporcionou uma redução de 39 para 18 fatores que podem ser utilizados na predição da TR.

Foram desenvolvidos, com sucesso, quatro modelos neurais que podem ser utilizados para predizer a TR. O primeiro deles considera como entrada, dados sobre os 39 fatores considerados como preditivos da TR. O segundo modelo consistiu na associação entre a Análise de Componentes Principais e as Redes Neurais Artificiais. Nesse caso, foram considerados como variáveis de entrada, os dados referentes aos 18 fatores indicados como mais importantes segundo a Análise de Componentes Principais. Os outros dois modelos são modificações do segundo, em que o conjunto de variáveis de entrada foi modificado de acordo com aspectos práticos (análise clínica dos pacientes) observados no cotidiano. *A*

priori, todos os modelos são eficientes e podem ser utilizados para tal tarefa. No entanto, necessitam de posterior validação com dados de outras populações para que sua confiabilidade seja testada.

Em relação à disseminação dos resultados, a partir deste estudo foram apresentados e publicados em anais de congresso os seguintes trabalhos:

1. Romano, A. V. C. ; MARTINS, T. D. ; Maciel Filho, R. ; Paula, E. V. ; Annichino-Bizzacchi, J. M. . Artificial Neural Network for Prediction of Venous Thrombosis Recurrence. In: ASH - 58th Annual Meeting and Exposition, 2016, San Diego. Proceedings of ASH - 58th Annual Meeting and Exposition. Blood, 2016. v. 128. p. 3771.

2. MARTINS, T. D.; Romano, A. V. C. ; Maciel Filho, R. ; Annichino-Bizzacchi, J. M. . Artifical neural network for prediction of venous thrombosis recurrence. In: HEMO 2016 - Congresso Brasileiro de Hematologia, Hemoterapia e Terapia Celular, 2016, Florianópolis. Revista brasileira de Hematologia e Hemoterapia. Ribeirão Preto: Associação Brasileira de Hematologia, Hemoterapia e Terapia Celular, 2016. v. 38. p. 106-107.

Além disso, até o presente momento foram submetidos para publicação seguintes artigos completos a revistas indexadas:

1. T.D. Martins, A.V.C. Romano, J. M. Annichino-Bizzacchi, R. Maciel Filho. Principal Component Analysis on Recurrent Venous Thromboembolism. Thrombosis Research.

2. T.D. Martins, A.V.C. Romano, J. M. Annichino-Bizzacchi, R. Maciel Filho. Artificial Neural Networks for Prediction of Recurrent Venous Thromboembolism. Artificial Intelligence in Medicine.

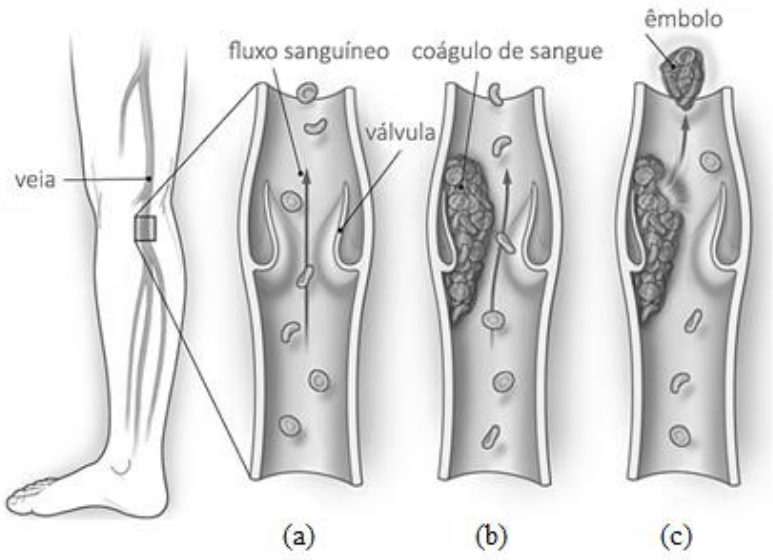
2 TROMBOSE VENOSA PROFUNDA

O principal foco deste trabalho é a TR de membros inferiores e do sistema nervoso central. Sendo assim, os tópicos desse Capítulo serão dedicados aos conceitos básicos para a compreensão do leitor sobre a coagulação sanguínea, a trombose, seus fatores de risco e recorrência.

2.1 Introdução

A trombose venosa consiste na formação de um trombo no interior de uma veia, geralmente nos membros inferiores, ou em outros sítios, como veias cerebrais, abdominais, dos membros superiores, entre outros. A Figura 2.1 ilustra a formação de um trombo em uma veia de um membro inferior. A Figura 2.1a mostra o fluxo sanguíneo de um indivíduo saudável. Na Figura 2.1b um trombo formado na parede venosa é ilustrado. Nota-se que o fluxo sanguínedo fica comprometido nesse caso. A Figura 2.1c mostra o desprendimento de parte do trombo, originando um embolo que pode dar origem a embolia pulmonar.

Figura 2.1 – Formação do trombo em uma veia de um membro inferior.



FONTE: Adaptado de (Veinguide.Com, 2003).

Quando a formação ocorre em uma veia do sistema profundo, é denominada TVP. Uma complicação da sua fase mais aguda é a EP, que consiste no desprendimento e deslocamento de um trombo (ou parte dele) para a árvore arterial pulmonar (Figura 2.1c). A trombose que ocorre nas veias cerebrais é denominada TVC.

A maioria dos casos de TVP ocorre nas veias dos membros inferiores, devido à sua posição que propicia a estase sanguínea, ocasionando um acúmulo de plaquetas e de fatores de coagulação, que são ativados e iniciam a formação do trombo. A cabeça do trombo, como é denominada, se adere à parede venosa e aumenta de tamanho na direção do fluxo sanguíneo, originando a cauda do trombo, que pode obstruir a veia, ou se desprender e se deslocar para o pulmão. Existem complicações tardias que também podem ocorrer, tais como a síndrome pós-trombótica, com o aparecimento de varizes secundárias, formação de úlceras, edemas, e dermatite ocre (Zago *et al.*, 2004).

2.2 A Coagulação do Sangue em um Indivíduo Saudável

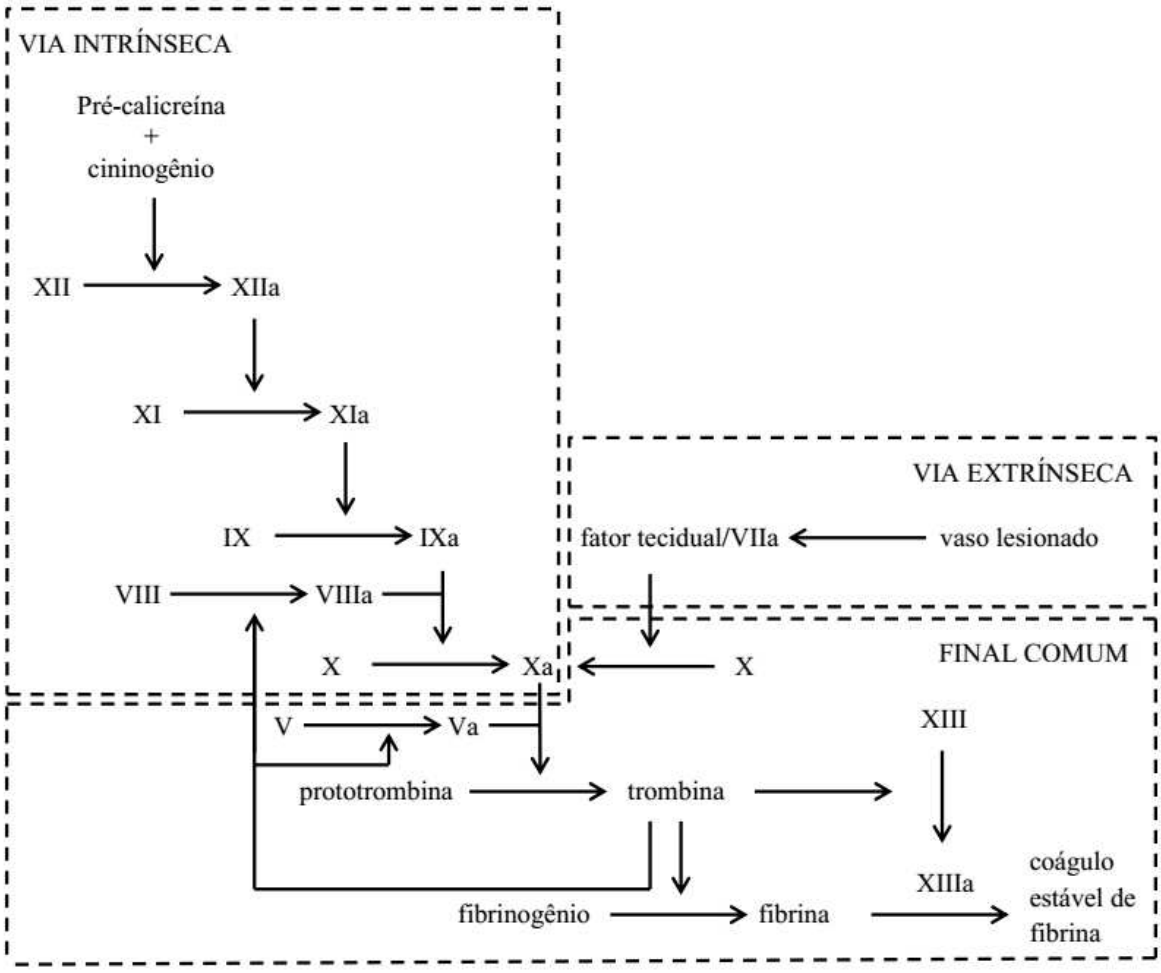
A coagulação do sangue é um conjunto de reações químicas que terminam na formação de polímeros de fibrina insolúvel, o coágulo. Em um indivíduo saudável, o sangue não coagula na corrente sanguínea devido ao equilíbrio das reações de ativação e inibição da coagulação. A TVP é um distúrbio que acontece quando algo provoca o desequilíbrio das reações químicas dessa sequência, gerando um trombo dentro do vaso (Zago *et al.*, 2004).

2.2.1 O Modelo Clássico da Coagulação

O modelo clássico da coagulação do sangue é chamado de cascata da coagulação e foi proposto em 1964, por Macfarlane, Davie e Ratnoff (Davie; Ratnoff, 1964; Macfarlane, 1964). Ele consiste na ativação sequencial de pró-enzimas por proteases do plasma resultando na formação da trombina, que é o catalisador da conversão da molécula de fibrinogênio em fibrina. As reações de formação do coágulo consistem na atuação de fatores pró e anticoagulantes, juntamente com os fatores pró e antifibrinolíticos (Zago *et al.*, 2004). Além dos fatores, existem vários cofatores que atuam nessa sequência de reações. A Figura 2.2 ilustra essa sequência de reações.

Os autores sugeriram a divisão do processo de coagulação em duas vias: extrínseca e intrínseca. A diferença entre as duas rotas é basicamente o iniciador de cada reação. Enquanto a via intrínseca é iniciada por substâncias que normalmente estão no sangue, a via extrínseca é iniciada devido à combinação de componentes presentes na circulação sanguínea com substâncias que normalmente não estão. Ambas as sequências terminam na mesma reação: a transformação do fator X em Xa. Em todo o texto o sufixo 'a' identificará um fator que foi ativado.

Figura 2.2 – Esquema da cascata da coagulação proposto por Macfarlane (1964).



A cascata da coagulação segue a seguinte sequência (Zago *et al.*, 2004): quando um vaso sanguíneo é danificado e o sangue é exposto a substâncias que normalmente não estão presentes na corrente sanguínea, a coagulação é iniciada. Assim, o fator tecidual (FT)

(ou tromboplastina, presente nas células subendoteliais), se liga ao fator VII, gerando o complexo FT-VIIa. Esse, por sua vez, transforma os fatores IX em IXa e X em Xa. Além disso, uma pequena concentração de trombina é formada como consequência dessa etapa.

Na via intrínseca, a formação de um complexo entre pré-calicreína, cininogênio de alto peso molecular e o fator XII, gera o fator XIIa, que converte o fator XI em XIa, que transforma o fator IX em IXa. O fator IXa, juntamente com o fator VIIIa, transforma o fator X em Xa. Por fim, a geração da trombina se dá a partir do fator Xa, que atua com o fator Va, transformando a protrombina (fator II) em trombina (fator IIa), que por sua vez converte o fribrinogênio em monômeros de fibrina. Além disso, a trombina também ativa o fator XIII, que estabiliza os monômeros de fibrina, formando polímeros que compõem os coágulos.

É importante salientar que a reação inicial de transformação dos fatores V e VIII, nos fatores Va e VIIIa, ocorre a partir da presença de traços de trombina gerada pela via extrínseca.

2.2.2 O Modelo Coagulação Revisado

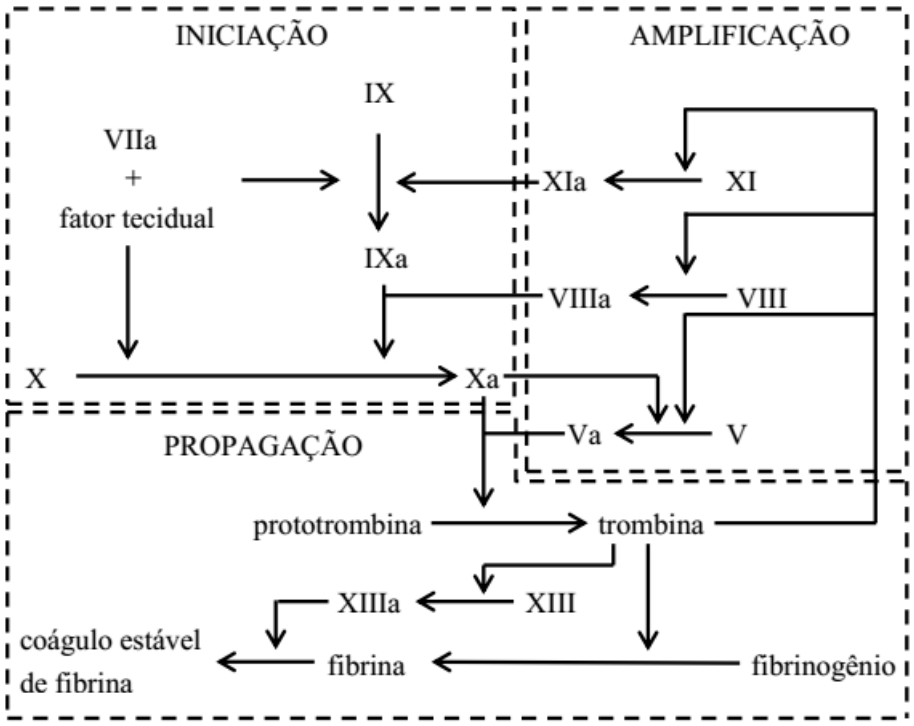
Apesar de ser aceito durante quase cinquenta anos, o modelo clássico da coagulação não explica por que indivíduos que possuem deficiência nos fatores VIII e IX apresentam graves manifestações hemorrágicas, e por que indivíduos com deficiência em fator XII, pré-calicreína ou cininogênio de alto peso molecular não apresentam sangramento (Zago *et al.*, 2004; Ferreira *et al.*, 2010). Além disso, diversos avanços mostraram que o modelo de uma cascata que age por duas vias independentes não ocorreria *in vivo*.

Assim, uma revisão do modelo clássico foi proposta por Hoffman e Monroe III (2001). Essa modelo se baseia no fato de que a formação do coágulo exige que os fatores ativados estejam sempre presentes no sítio da lesão. Além disso, entendeu-se que o início do processo se dá com a exposição do FT na corrente sanguínea, e que as reações de ativação ocorrem em células contendo fosfolipídeos em sua membrana, e se dividiu a sequência de reações em três etapas: inicialização, amplificação e propagação (Ferreira *et al.*, 2010), conforme ilustra a Figura 2.3.

Na fase de iniciação, as células subendoteliais que expressam o FT em sua superfície são expostas aos componentes do sangue no sítio da lesão. O FT também pode ser expresso nos monócitos ou em micropartículas circulantes. O FT transforma o fator VII em VIIa e se liga a ele formando um complexo que é capaz de gerar pequenas quantidades dos fatores IXa e Xa. O fator V é transformado em Va pelo fator Xa e por proteases não coagulantes. Assim, juntos, os fatores Xa e Va formam o complexo protrombinase na superfície da célula que expressa o FT. Nessa estapa, o complexo protrombinase só responde pela formação de pequenas quantidades de trombina. Quando há um dano vascular grave, o processo de coagulação segue para a próxima fase.

A fase de amplificação consiste na transformação dos fatores V, VIII e XI em Va, VIIIa e XIa, respectivamente. Nessa fase, a trombina gerada na fase de iniciação ativa as plaquetas presentes no sangue e também intermedia a ativação dos fatores supracitados na superfície das plaquetas.

Figura 2.3 – Esquema da cascata da coagulação revisado.



A fase de propagação consiste na estabilização do coágulo de fibrina no sítio da lesão. Para isso, a quantidade de plaquetas no local lesionado aumenta drasticamente. O fator IXa, gerado na fase inicial e pelo fator XIa, se liga ao fator VIIIa na superfície das plaquetas, e forma o complexo tenase que é responsável pela produção de maiores quantidades do fator Xa. Em seguida, o fator Va ligado na plaqueta se liga ao fator Xa, resultando na formação do complexo protrombinase que gera grandes quantidades de trombina, que forma a fibrina e ativa o fator XIII, que estabiliza o coágulo (Ferreira *et al.*, 2010).

Em todas as reações de ativação, diversos cofatores são necessários dentre eles os íons cálcio, e que agem promovendo a ligação da protrombina e dos fatores VII, IX e X aos fosfolípides plaquetários.

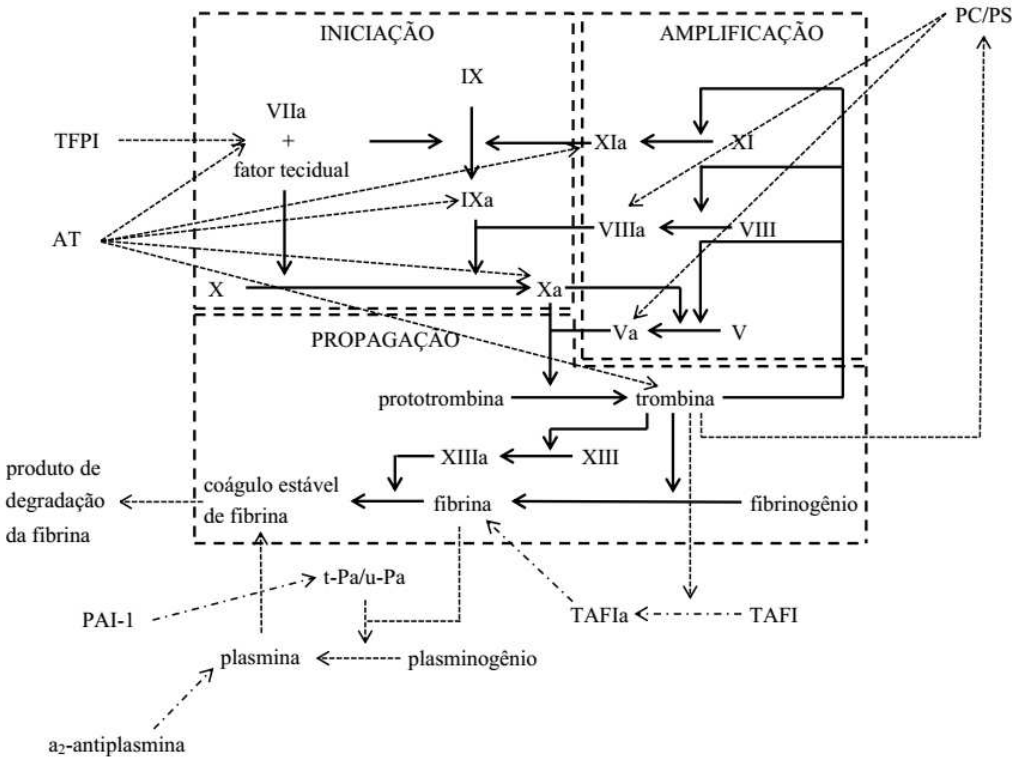
2.2.3 Mecanismos Reguladores da Coagulação

Uma vez que o coágulo de fibrina está formado e estável no local lesionado, é necessário um mecanismo que faça com que a ativação do sistema seja contida a fim de não haver o bloqueio do vaso sanguíneo. Assim, mecanismos anticoagulantes e de dissolução do coágulo (fibrinólise) ocorrem de modo a equilibrar as reações de coagulação.

Existem quatro anticoagulantes naturais que atuam no controle das reações de coagulação: o inibidor da via do fator tecidual (TFPI), a proteína C (PC), a proteína S (PS), e a antitrombina (AT). Cada um atua na inibição de um determinado fator coagulante, conforme ilustrado na Figura 2.4. Nessa figura, as linhas contínuas representam as reações de coagulação e as tracejadas, os mecanismos de anticoagulação e fibrinólise.

O TFPI é uma proteína produzida pelas células endoteliais e inibe a geração de fator Xa se ligando a ele e posteriormente ao complexo FT-VIIa, formando um complexo quaternário e limitando as reações de coagulação. A PC, quando ligada ao endotélio, é ativada pela trombina que está ligada à proteína transmembrana trombomodulina (IIa-TM). O poder inibitório da PC é potencializado pela PS, que atua como um cofator não enzimático na clivagem dos fatores Va e VIIIa, inativando-os. A AT inibe a atividade da trombina e dos fatores IXa, Xa, XIa e XIIa, e acelera a dissociação do complexo FT-VIIa. Seu efeito é potencializado por ação de substâncias presentes no endotélio, como o heparan sulfato (Zago *et al.*, 2004; Ferreira *et al.*, 2010).

Figura 2.4 – Esquema da cascata da coagulação incluindo os mecanismos reguladores.



Outro mecanismo regulador da coagulação é a dissolução dos polímeros de fibrina, denominado fibrinólise, também ilustrado na Figura 2.4. Nessa etapa, o ativador do plasminogênio do tipo tecidual (t-PA) e o ativador do plasminogênio do tipo uroquinase (u-PA) hidrolisam o plasminogênio, gerando a plasmina, que degrada a fibrina. Essa reação é altamente específica, visto que o t-Pa tem baixa afinidade pelo plasminôgenio na ausência de fibrina. Por outro lado, também existe um mecanismo de inibição da fibrinólise (representado na Figura 2.4 com linhas traço-ponto): o inibidor do ativador do plasminogênio (PAI-1) evita a formação da plasmina, e a α_2-antiplasmina impede a ação da plasmina. Por fim, há também outro inibidor da fibrinólise, denominado inibidor da fibrinólise ativado pela trombina (TAFI), que é ativado pelo complexo IIa-TM (TAFIa) e atua inibindo a geração de plamina.

2.3 Incidência, Tratamento e Recorrência da TVP

A incidência de TVP no Brasil não é ampla e sistematicamente investigada. Ao que se sabe, a última estatística vem de dados do SUS entre os anos de 2008 e 2010 que indicaram 85772 internações decorrentes de TVP, sendo que a taxa de mortalidade foi de 2,38 %. Sabe-se, também, que a TVP causa mais mortes que o HIV, câncer de próstata, câncer de mama e acidentes automobilísticos juntos. Por fim, que a EP causa 1 morte a cada 6 indivíduos, em até 3 meses após sua ocorrência (Ibope, 2010).

Uma vez diagnosticado, o paciente com TVP precisa ser tratado o mais rápido possível. O tratamento da trombose possui diversos objetivos: i) prevenir a recorrência da doença, ii) prevenir EP naqueles que possuem apenas TVP, iii) prevenir que o trombo avance no local que foi gerado ou que tromboses ocorram em outros vasos e, iv) prevenir síndrome pós-trombótica (Hirsh; Hoak, 1996). A administração de anticoagulantes é o principal tratamento e estão disponíveis diversas opções capazes de inibir a coagulação do sangue em diferentes vias.

A heparina inibe principalmente a trombina e o fator Xa, catalisando a reação desses fatores com a AT. A warfarina é um anticoagulante definido como antagonista da vitamina K, e compete com a epóxiredutase pela redução da vitamina K. A vitamina K reduzida é responsável pela γ-carboxilação dos fatores II, VII, IX, e Xa. A rivaroxabana, a apixabana e a edoxabana são inibidores orais diretos do fator Xa e, a dabigatrana é um inibidor oral direto da trombina. Além disso, há diversos novos anticoagulantes sendo desenvolvidos e testados (Fernandes *et al.*, 2016). O *American College of Chest Physicians* (ACCP) *Guidelines* (Kearon et al., 2016), sugere a que o tratamento do TEV deve durar no mínimo 3 meses. Após esse período, o paciente deve passar por nova avaliação clínica e laboratorial e assim é tomada a decisão de continuar ou não o tratamento. Uma complicação frequentemente encontrada é a recorrência da trombose.

A TR é uma complicação multifatorial, e ainda não completamente compreendida, que atinge de 0,6 % a 5 % dos pacientes nos primeiros 90 dias e de 13 % a 30 % nos primeiros 5 anos (Farzamnia *et al.*, 2011). Assim, um grande desafio é prever quais os pacientes que desenvolverão ou não a TR. É sabido que o tratamento anticoagulante previne a

ocorrência de TR (Franco *et al.*, 2017), mas determinar apropriadamente o tempo de duração do tratamento deve levar em conta dois aspectos: o risco de TR *versus* o risco de sangramento do uso do anticoagulante. Um paciente com alto risco de TR, usualmente mantém maior tempo de tratamento com o anticoagulante; por outro lado, pacientes com alto risco de sangramento precisam ser medicados com cautela. A busca por uma ferramenta que determine a probabilidade dos dois tipos de risco com precisão ainda é um assunto em aberto.

Em geral, pacientes que apresentam TR espontânea devem ser tratados com anticoagulantes indefinidamente. No entanto, esses pacientes acabam expostos a um alto risco de sangramento. Diferentes métodos matemáticos foram propostos para se avaliar o risco hemorrágico, mas tais modelos não são precisos e não foram validados na TVP (Riva *et al.*, 2014; Klok *et al.*, 2016). Por outro lado, pacientes não diagnosticados como propensos à TR correm risco de apresentar um novo episódio trombótico, assim como de EP (Kyrle, 2016).

2.4 Fatores de Risco da Trombose Venosa Profunda Recorrente

A tríade de Virchow (Bagot; Arya, 2008) postula que três são as principais causas para a trombose: alterações na coagulabilidade do sangue, alterações nas paredes dos vasos sanguíneos e estase venosa. Segundo Kyrle e Eichinger (2009), diversos trabalhos mostram que o risco de trombose e de TR aumenta por alterações nessas vias. Sabe-se que o antecedente de um episódio trombótico é o principal fator de risco para a TR. Assim, diversas características da manifestação do primeiro episódio trombótico, assim como do seu tratamento são importantes aspectos a serem levados em conta.

Na literatura podem ser encontrados diversos artigos de revisão que enumeram vários fatores de risco e preditivos para a TR (Zhu *et al.*, 2009; Fahrni *et al.*, 2015; Rosendaal, 2016). Neste trabalho, além dos fatores de risco já definidos, também foram considerados diversos fatores que podem ser candidatos a afetar esse risco. Nessa Seção, uma breve descrição de todos os fatores será apresentada, bem como uma revisão da sua influência nos processos trombóticos. Este conhecimento é importante para a seleção das variáveis que poderão compor o modelo matemático, bem como na discussão dos resultados obtidos.

2.4.1 Sexo, Idade e IMC

As características dos indivíduos que apresentam um primeiro episódio trombótico são importantes para o cálculo do risco de TR. Assim, sexo, idade e o índice de massa corporal (IMC) são fatores relacionados de alguma forma com a TR e, por esse motivo, foram inclusos neste trabalho.

O sexo do paciente como fator para a TR é tratado em diversos trabalhos. Apesar de importante, esse fator só foi levado em consideração a partir de 2004, no estudo de Kyrle *et al.* (2004). Em geral, há um consenso de que homens têm maior risco de recorrência do que as mulheres (Mcrae *et al.*, 2006; Eichinger *et al.*, 2008; Christiansen *et al.*, 2010; Farzamnia *et al.*, 2011; Galanaud *et al.*, 2014). Segundo alguns estudos, isso ocorre porque as mulheres apresentam a primeira TVP quando estão expostas a fatores de risco, tais como gravidez e contraceptivos orais, ou seja, quando são jovens. Em geral, a primeira TVP em homens se dá em idades mais avançadas e por esse motivo, a recorrência em homens seria mais frequente (Lijfering *et al.*, 2009; Lijfering *et al.*, 2010).

Diversos estudos também associam idade avançada com alto risco de TVP/TVC e TR e indicam que apenas 0,01 % da população possui eventos trombóticos antes de atingir 40 anos. Por outro lado, essa proporção aumenta 70 vezes para aqueles entre 45 e 55 anos. Além disso, a probabilidade de morte aumenta com o avanço da idade. (Silverstein *et al.*, 1998; Tsai *et al.*, 2002; Fahrni *et al.*, 2015). Apesar disso, Hansson *et al.* (2000) e Farzamnia *et al.* (2011) não encontraram associação entre a idade e TR. Por outro lado, White *et al.* (1998) reportaram que a recorrência é mais comum em pacientes jovens e, Heit *et al.* (2000), Eichinger *et al.* (2007) e Galanaud *et al.* (2014) observaram uma associação entre idade avançada e o maior risco de recorrência.

A obesidade é outra característica do paciente a ser levada em conta. Diversos estudos relacionam a obesidade de alguma forma com um maior risco de TR. Alguns trabalhos relacionam a medida da circunferência abdominal (Cushman *et al.*, 2016), e outros o índice de massa corporal (IMC). Heit *et al.* (2000) reportaram que a TR é 24 % mais frequente em indivíduos com maior IMC. Eichinger *et al.* (2008) também observaram uma relação entre o aumento do IMC e o maior risco da TR. Por outro lado, Farzamnia *et al.*

(2011) não encontraram relação entre o IMC e a TR. Vučković *et al.* (2017) também reportaram que a relação entre IMC e a TR, somente está presente em pacientes do sexo feminino.

2.4.2 Natureza da Trombose e Concentração de D-dímero

O tratamento anticoagulante e o risco de TR são tratados de forma diferente dependendo da natureza da TVP. Sabe-se que o primeiro episódio trombótico pode ocorrer de forma espontânea ou ser provocado por fatores de risco momentâneos. Quando é provocada, usualmente o risco de TR é baixo quando o fator desencadeante não existe mais. Assim, tais pacientes são tratados durante 3 meses e depois o tratamento é descontinuado. Para o caso das TVPs espontâneas, a continuidade da anticoagulação depende de fatores clínicos e laboratoriais após a finalização do tratamento inicial padrão.

Na literatura encontram-se diferentes opiniões sobre esse tópico. Iorio *et al.* (2010) reportaram que o risco de TR em pacientes que tiveram uma primeira TVP espontânea é muito maior em relação àqueles que tiveram uma TVP provocada. Por outro lado, no estudo de Cosmi *et al.* (2011) foi observado que pacientes que tiveram uma primeira TVP provocada, e que apresentaram altos níveis de D-dímero durante o tratamento anticoagulante ou um mês após o final do tratamento possuem de 7 % a 11 % de risco de TR.

O nível de D-dímero é, atualmente, um dos principais fatores a ser acompanhado em um paciente com antecedente de TEV. O D-dímero é um produto de degradação da fibrina e está presente em indivíduos saudáveis (nível de até 500 $ng.mL^{-1}$). No entanto, quando ocorre um TEV, seus níveis aumentam significativamente. A relação entre o alto nível de D-dímero e TR é apresentada em diversos estudos. Eichinger *et al.* (2003) reportaram que os pacientes que tiveram TR em geral apresentavam níveis de D-dímero significativamente maiores em relação aos que não apresentavam a recorrência. Além disso, os pacientes que apresentavam nível de D-dímero abaixo de 250 ng/mL foram os que menos tiveram TR. Palareti *et al.* (2014) acompanharam a evolução dos níveis de D-dímero 15, 25, 55 e 85 dias após o término da anticoagulação. Os resultados mostraram que altos níveis de D-dímero estão associados à TR. Em um estudo recente, Bjøri *et al.* (2017) reportaram que pacientes com nível de D-

dímero alto no momento do diagnóstico da primeira TVP (acima de 1500 ng/mL) possuem maior risco de TR do que aqueles com baixo nível de D-dímero.

2.4.3 Localização do Trombo

A TVP de membros inferiores pode ocorrer do lado direito, esquerdo, ou em ambos os lados. Além disso, ela normalmente é dividida conforme sua localização. É dita proximal quando ocorre na veia ilíaca e/ou femoral e/ou poplítea com ou sem trombose em veias da perna. É dita distal quando acomete somente as veias da perna. É importante se definir a localização da trombose, pois indivíduos com trombose proximal possuem risco maior de TR do que aqueles com trombose distal.

De acordo com Galanaud *et al.* (2014), ainda há controvérsias sobre os riscos de uma TVP distal ou proximal para a TR. Para auxiliar nesse debate, em seu estudo, os autores acompanharam 490 pacientes com uma primeira TVP distal e 259 com uma primeira TVP proximal ao longo de 3 anos. Os resultados mostraram que aqueles indivíduos com TVP proximal tiveram mais propensão à TR (5,2 %) do que aqueles com TVP distal (3,7 %). Resultados semelhantes foram reportados previamente, por Boutitie *et al.* (2011) e Hansson *et al.* (2000). O estudo de Hansson *et al.* (2000) também mostrou que pacientes que tiveram a primeira TVP do lado esquerdo são mais propensos à recorrência.

A EP é uma manifestação mais grave do TEV, sendo muitas vezes uma doença fatal. Por esse motivo, é necessário determinar se aqueles pacientes que tiveram uma EP possuem maior risco de TR em relação aos outros indivíduos. Prandoni *et al.* (2007) reportaram que pacientes com uma primeira EP tem maior risco de recorrência em relação à pacientes com uma primeira TVP.

O trombo residual também é um fator que pode estar associado com a TR. O estudo de Siragusa *et al.* (2008) mostrou que após a primeira TVP a recorrência ocorreu em 27,2 % daqueles que pararam com o tratamento anticoagulante, e 19,3 % daqueles que continuaram com o tratamento. Carrier *et al.* (2011) também mostraram que o trombo residual pode estar associado com maior risco de TR em pacientes com uma primeira TVP provocada ou espontânea. Por outro lado, Cosmi *et al.* (2011) reportaram que o trombo residual não está associado a um maior risco de TR.

2.4.4 Tempo do Tratamento de Anticoagulação

Como mencionado na Seção 2.3, o tempo do tratamento de anticoagulação é um dos fatores mais importantes no tratamento e na prevenção da TR. Esse tópico ainda gera controvérsias e diferentes pontos de vista são encontrados na literatura. Todos os pacientes são tratados com anticoagulantes durante 3 meses e após esse período a decisão de continuar ou não com o tratamento depende de cada caso (Kearon *et al.*, 2016). A continuidade ou não do tratamento após os 3 meses da administração do anticoagulante é alvo de muitos debates no meio científico. Eichinger *et al.* (2008) reportaram que não há relação entre o tempo do tratamento de anticoagulação e a TR. Boutitie *et al.* (2011) descreveram que a TR é menos frequente em pacientes tratados por 3 a 6 meses, em relação àqueles que foram tratados ao longo de 1-1,5 meses e 3 meses, respectivamente. Nieto Rodríguez e Ramírez Luna (2017) apresentaram diversas evidências que sugerem a redução tempo de tratamento anticoagulante para o mínimo possível, especialmente para pacientes que tiveram uma TVP provocada por fatores de risco temporários, isolada, ou com alto risco de sangramento.

Neste trabalho, o tempo do tratamento de anticoagulação será incluído como variável independente, especialmente devido aos resultados reportados por Agnelli *et al.* (2013), que demonstraram que o prolongamento do tempo de anticoagulação com apixabana reduziu o risco de recorrência sem aumentar o risco de sangramento. Resultados semelhantes também foram encontrados em The EINSTEIN Investigators (2010) e em Hansson *et al.* (2000).

2.4.5 Fatores de Risco Genéticos

As trombofilias hereditárias, como o fator V Leiden e a mutação G20210A no gene protrombina são comumente associadas com a TVP, mas a sua relação com a TR ainda não é completamente elucidada. Além disso, o nível elevado de fator VIII é outra variável associada à TVP.

Eichinger *et al.* (2003) reportaram que esses três fatores não influenciam fortemente a TR, quando comparados com os níveis de D-dímero. Eichinger *et al.* (2007) referiram posteriormente que indivíduos que possuem altos níveis de fator VIII três semanas

após o término do tratamento anticoagulante têm maior risco de TR. No entanto, não encontraram uma associação com o fator V Leiden e a mutação G20210A no gene prototrombina. Eichinger *et al.* (2008) e Méan *et al.* (2017) também reportaram resultados semelhantes.

Apesar dos resultados encontrados na literatura, dentro de um contexto multifatorial esses fatores podem influenciar, mesmo que de forma leve, na incidência de TR. Além disso, a população avaliada neste trabalho é diferente da população pesquisada nos trabalhos supracitados. Assim, eles foram incluídos na análise.

2.4.6 Anticoagulantes Naturais e Síndrome do Anticorpo Fosfolipídeo

O baixo nível das proteínas anticoagulantes (PC, PS e AT) leva à hipercoagulabilidade e predispõe a eventos trombóticos (Heeb *et al.*, 1994; Pabinger; Schneider, 1996; Esmon, 2000; Kim *et al.*, 2014). Rosendaal e Reitsma (2009) encontraram uma forte associação entre TVP e a baixa concentração dessas proteínas, mas afirmaram que a relação com a TR não é necessariamente verdadeira e que outros estudos precisam ser realizados. De Stefano *et al.* (2006) acompanharam pacientes que possuíam deficiência nos níveis dessas proteínas no sangue e reportaram que esses indivíduos possuem risco elevado para TR.

A síndrome do anticorpo antifosfolipídeo (SAF) é uma doença que afeta a coagulação do sangue e é caracterizada por tromboses venosas e arteriais. Indivíduos com SAF tem um risco muito elevado de TVP e TR em relação à população em geral. Pengo *et al.* (2010) reportaram que indivíduos com SAF possuem maior risco de TR em relação àqueles que não apresentam a síndrome, mesmo durante o tratamento anticoagulante. Mais recentemente, Medina *et al.* (2017) e Comarmond *et al.* (2017) reportaram resultados semelhantes.

2.4.7 Lipídios e Glicemia

Dentre os fatores de risco para a TR também podem se encontrar componentes metabólicos do sangue. Alguns estudos abordaram a importância dos lipídios para a TVP, mas

os resultados são contraditórios. (Deguchi *et al.*, 2005; Everett *et al.*, 2009). Morelli *et al.* (2017) avaliaram a associação entre a TR e colesterol total, LDL, HDL, triglicérides e apolipoproteínas A1 e B. Os resultados mostraram que não há influência desses lipídeos e sobre a TR. Por outro lado, Eichinger *et al.* (2007) reportaram que altos níveis de HDL e apoliproteína A1 estão associados com maior risco de TR. Ainda não se sabe se o nível de glicose no sangue tem relação com a TR. Porém, há estudos que relacionam o seu alto nível com a hipercoagulabilidade em pacientes com diabetes mellitus (Grant, 2007; Kim *et al.*, 2014).

2.4.8 Variáveis Hematológicas

Diversos estudos associam os níveis das células vermelhas, brancas e das plaquetas com a TVP ou com a TR. Podem ser encontrados estudos que relatam a influência do número de células vermelhas (Alt *et al.*, 2002; Vayá; Suescun, 2013; Litvinov; Weisel, 2017) e seu tamanho (Rezende *et al.*, 2013; Bucciarelli *et al.*, 2015), nível de hemoglobina (Brækkan *et al.*, 2010), número de leucócitos (Carobbio *et al.*, 2007; De Stefano *et al.*, 2008; Rezende *et al.*, 2013), e número de plaquetas e seu volume médio (Rupa-Matysek *et al.*, 2014). Supõe-se que a influência desses fatores se deve principalmente à alterações na viscosidade do sangue, e da sua coagulabilidade (Wells; Merrill, 1962; Dintenfass, 1964; Yu *et al.*, 2011).

2.4.9 Fatores de Risco Adquiridos

Os fatores de risco transitórios são aqueles que uma vez presentes podem provocar uma TVP ou TR, mas que quando desaparecem, o risco de trombose também passa a não existir. Tais variáveis envolvem gravidez e puerpério (Heit *et al.*, 2000; Farzamnia *et al.*, 2011; Barillari *et al.*, 2016), câncer (Hansson *et al.*, 2000; Heit *et al.*, 2000; Farzamnia *et al.*, 2011), hipertensão (Huang *et al.*, 2016), diabetes (Hess; Grant, 2011; Chung *et al.*, 2015), dislipidemia (García *et al.*, 2014), insuficiência hepática e renal (Shlipak *et al.*, 2003; Anthony Lizarraga *et al.*, 2010; Aggarwal *et al.*, 2014), uso de hormônios estrogênicos (Galanaud *et al.*, 2014), assim como tabagismo (Farzamnia *et al.*, 2011). Todos esses fatores foram inclusos neste trabalho visando se obter um modelo amplo para a predição da TR.

2.5 Cálculo do Risco de Recorrência de Trombose Venosa Profunda

A tomada de decisão para prosseguir ou não com o tratamento anticoagulante para além de três meses não é uma tarefa fácil. Até onde se sabe, a decisão ainda é tomada de forma empírica, com base na análise dos resultados de exames e sintomas do paciente ao longo e ao fim do tratamento. Uma das formas de se abordar esse problema é conhecendo-se a fenomenologia da formação do coágulo e suas complicações, equacionando todos os parâmetros usando expressões matemáticas.

Realizar essa tarefa não é fácil, visto que o sistema de coagulação sanguíneo depende de diferentes fatores, dentre eles: o conjunto das reações que formam o coágulo e aquelas que fazem o controle da coagulação, aspectos do fluxo sanguíneo, componentes do sangue, fatores genéticos, etc. Um passo inicial já foi dado nessa direção: podem ser encontrados na literatura alguns estudos que modelam matematicamente as reações da cascata de coagulação do sangue. Cada um possui suas vantagens, mas muito ainda precisa ser feito para a obtenção de um modelo completo (Zhu, 2007; Xu *et al.*, 2010; Lacroix, 2012; Leiderman; Fogelson, 2014).

Por outro lado, alguns métodos matemáticos empíricos foram propostos para se calcular o risco de TR baseados em dados clínicos após o primeiro TEV. Todos foram desenvolvidos a partir de métodos estatísticos aplicados a determinadas populações e cada um leva em conta diferentes fatores para o cálculo. Nas próximas seções, os aspectos teóricos de cada método serão apresentados, bem como suas vantagens e desvantagens.

2.5.1 HERDOO2 *Score*

O primeiro modelo matemático proposto para se calcular o risco de TR foi o de Rodger *et al.* (2008) e é denominado HERDOO2. Nesse trabalho, os autores visavam obter um modelo para identificar pacientes que pertencem a um grupo cuja TR é pouco provável: aqueles com baixo risco de recorrência (< 3 % ao ano). O estudo realizado pelos autores incluiu 646 pacientes de diferentes países que apresentaram uma primeira TVP proximal ou EP espontânea. Foram coletadas informações a respeito de 69 variáveis e os fatores preditivos da TR foram determinados utilizando-se um método de regressão logística condicional.

Os resultados obtidos pelos autores mostraram que 91 pacientes apresentaram episódios de TR e que não foi possível identificar padrões para pacientes do sexo masculino. Por outro lado, foi possível formular uma regra para identificar pacientes do sexo feminino que estivessem no grupo de baixo risco. Segundo os autores, mulheres que apresentam 0 ou 1 dos seguintes parâmetros não necessitam de terapia prolongada:

1. Nível de D-dímero acima de 250 μg/L;
2. Índice de massa corporal acima de 30 kg/m²;
3. Idade acima de 65 anos;
4. Sintomas pós-trombose (hiperpigmentação, edema ou vermelhidão nas pernas).

Em um trabalho recente, Rodger *et al.* (2017) realizaram um estudo de validação do modelo proposto previamente, acompanhando 2785 pacientes de diferentes países, e que apresentavam as mesmas características do trabalho anterior. Os resultados comprovaram a eficácia do modelo, uma vez que apenas 17 mulheres (de um total de 631) classificadas no grupo de baixo risco desenvolveram TR.

Apesar dos bons resultados, pode-se citar dentre as principais desvantagens desse modelo: i) exclusão de pacientes do sexo masculino ii) incapacidade de predição do tempo ótimo de duração da anticoagulação, iii) exclusão de pacientes do grupo de alto risco, e iv) incidência de TR em mulheres acima de 50 anos foi maior que o previsto pelo modelo.

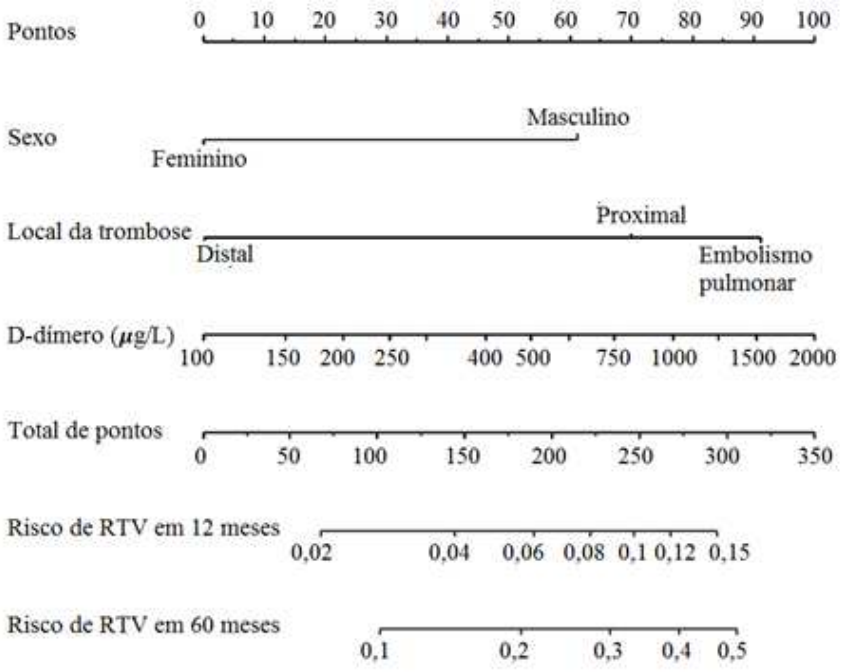
2.5.2 Vienna *Score*

Eichinger *et al.* (2010) propuseram o segundo modelo cujo objetivo foi calcular o risco de TR. Tal modelo foi denominado Vienna *score*. Em seu estudo, os autores acompanharam 929 pacientes de 4 centros hospitalares diferentes, e que apresentaram uma primeira TVP e EP espontâneas. Foram coletadas informações a respeito de 8 fatores e os fatores preditivos da TR foram determinados utilizando-se o método de regressão de Cox.

Os resultados mostraram que 176 pacientes apresentaram TR (correspondendo a 18,9 %) e foi reportado que os altos níveis de D-dímero, sexo masculino, TVP proximal, e EP estão associados com alto risco de TR. Com esses fatores, os autores desenvolveram um

nomograma que pode ser utilizado para o cálculo do risco de recorrência, apresentado na Figura 2.5.

Figura 2.5 – Nomograma do Vienna *score*.



Fonte: Adaptado de Eichinger *et al.* (2010).

A utilização do nomograma é simples e direta: com uma linha reta se relaciona o número de pontos e o sexo, nível de D-dímero 3 semanas após o fim do tratamento de anticoagulação e local da trombose. Em seguida, somam-se os pontos, obtendo o total de pontos. Por fim, relaciona-se o total de pontos com o risco de TR em 12 e 60 meses.

Apesar da simplicidade na utilização, esse método possui a desvantagem de ser aplicável apenas na terceira semana após o término do tratamento de anticoagulação. Por esse motivo, Eichinger *et al.* (2014) realizaram uma atualização que visava ampliar o intervalo de tempo em que o método poderia ser aplicado. Para isso, os autores levaram em consideração o grande poder preditivo da concentração de D-dímero, e acompanharam 553 pacientes

utilizando as medidas em 3 semanas, e 3, 9, 15 e 24 meses após a anticoagulação. Como resultado, os autores desenvolveram 4 nomogramas para o cálculo do risco de recorrência, sendo cada um aplicável para 3 semanas, 3 meses, 9 meses e 15 meses. Apesar da atualização promissora, o modelo não pode ser utilizado quando o paciente está há menos de 3 semanas, ou após 15 meses, do fim do tratamento anticoagulante.

Marcucci *et al.* (2015) realizaram um estudo independente que visava validar o Vienna *Score*. Em seu tempo, era o primeiro trabalho de validação de um *score* para TR. Para isso, os autores coletaram dados de 929 pacientes de quatro centros hospitalares diferentes, no período de 1992 e 2008. Os resultados mostraram que o modelo foi capaz de distinguir o risco de recorrência, mas que tende a subestimar o valor da probabilidade quando o paciente está 12 meses após a terapia de anticoagulação.

Por outro lado, Tritschler *et al.* (2015) visaram validar a atualização do Vienna *Score*, proposta por Eichinger *et al.* (2014), em pacientes acima de 65 anos. Nesse trabalho os autores coletaram dados de 156 pacientes de 9 centros hospitalares diferentes, entre 2009 e 2013. Os resultados mostraram que o método não foi capaz de diferenciar pacientes com alto risco de TR daqueles com baixo risco. Os autores sugeriram que isso possa ser devido ao alto valor da concentração de D-dímero, que naturalmente ocorre em pacientes com idades avançadas e sugerem que um modelo específico para essa faixa de idade seja formulado.

2.5.3 Dash *Score*

Tosetto *et al.* (2012) propuseram o mais recente método para o cálculo do risco de TR que se conhece. Tal modelo é denominado Dash *Score* e foi obtido de um estudo que acompanhou 1818 pacientes com uma primeira TVP espontânea. Foram coletadas informações sobre 6 fatores e os fatores preditivos da TR foram determinados utilizando-se o método de regressão de Cox.

Os resultados mostraram que 239 pacientes apresentaram TR (correspondendo a 13,1 %) e foi reportado que níveis elevados de D-dímero, sexo masculino, idade abaixo de 50 anos e uso de homônio (para mulheres) estão associados com alto risco de TR. Com os fatores relacionados ao alto risco de TR, os autores propuseram a seguinte pontuação para se calcular o risco de recorrência:

- D-dímero: +2
- Idade < 50 anos: +1
- Sexo masculino: +1
- Uso de hormônio: – 2

A soma da pontuação abaixo de um indica que a probabilidade anual de TR é abaixo de 5 % e o paciente pode interromper a anticoagulação com 3 meses de tratamento. Por outro lado, uma pontuação superior a um indica uma probabilidade anual acima de 5 % e sugere-se que esses pacientes continuem o tratamento. Em um estudo recente, Tosetto *et al.* (2017) realizaram a validação desse modelo em um estudo retrospectivo com 827 pacientes. Segundo os autores, o modelo pode ser utilizado na predição de TR, mas ainda precisa de refinamentos quando se trata de pacientes com idades acima de 65 anos.

2.6 Conclusões

Este Capítulo mostrou que há diversos fatores que influenciam a TR. Porém, ainda há muitas questões em aberto no que diz respeito aos fatores a ela relacionados. O número de fatores, bem como a complexidade dessas relações torna difícil e onerosa a obtenção de um modelo preditivo fenomenológico da TR. Apesar disso, diversos estudos já mostraram que todos os fatores aqui citados influenciam de alguma forma na recorrência da TVP/TVC e EP. Essa informação pode ser útil visto que, para o desenvolvimento de métodos de inteligência artificial, muitas vezes só é necessário se conhecer quais variáveis influenciam em determinado fenômeno (e não como se dá a relação entre elas).

3 REDES NEURAIS ARTIFICIAIS

Apesar do desenvolvimento e estudos de validação dos métodos de *scores*, ainda há muitas perguntas em aberto, especialmente no que diz respeito aos fatores preditivos. Além disso, ainda há a necessidade de aperfeiçoamentos no que se refere à prevenção da TR. Assim, neste trabalho foram propostos dois modelos matemáticos visando sua utilização no suporte à decisão clínica na prevenção de TR.

Na prática, o objetivo foi obter uma equação capaz de predizer a situação futura do paciente (retrombose ou não) baseada em fatores clínicos atuais. Assim, a fim de se obter essa equação, foram utilizadas as Redes Neurais Artificiais (RNAs), que tem ganhado bastante destaque recentemente. Nesse Capítulo serão abordados os conceitos básicos para a compreensão do funcionamento das RNAs, bem como suas aplicações recentes como ferramentas de suporte à decisão clínica.

3.1 Introdução

As RNAs são uma alternativa à modelagem fenomenológica de um fenômeno qualquer e a sua utilização vem ganhando espaço na ciência ao longo das últimas décadas. Isso se deve pricipalmente ao seu potencial de aprendizado e generalização, e também, à busca de novas tecnologias de modelagem que satisfaçam as necessidades de cada setor.

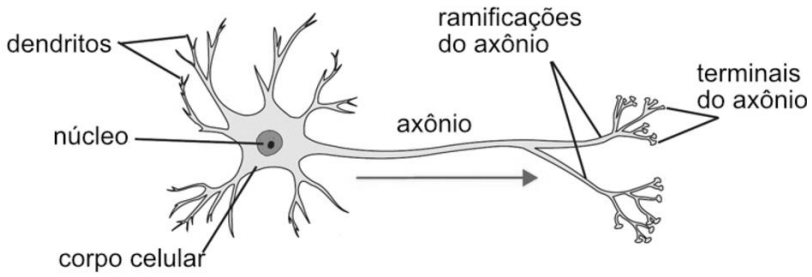
As RNAs foram desenvolvidas originalmente na década de 1940 por Warren McCulloch, do Massachusetts Institute of Technology, e por Walter Pitts, da Illinois University (Mcculloch; Pitts, 1943). Nesse estudo, os autores discorrem fazendo uma analogia entre os neurônios do cérebro humano e o funcionamento do processo eletrônico. A partir de então, diversos modelos de RNAs surgiram com o intuito de aperfeiçoar a técnica proposta originalmente e aplicá-la nas mais diversas áreas do conhecimento, como mostra a revisão publicada por Schmidhuber (2015).

3.2 O Neurônio Biológico

O funcionamento do neurônio biológico passou a ser mais bem entendido somente após da década de 40, quando este passou a ser considerado como uma unidade elementar do

sistema nervoso. Um neurônio biológico é composto por um corpo celular, pelo axônio, pelos dendritos e sinapses, conforme ilustrado na Figura 3.1.

Figura 3.1 – Ilustração de um neurônio biológico e suas principais características.



Fonte: Bezerra (2016).

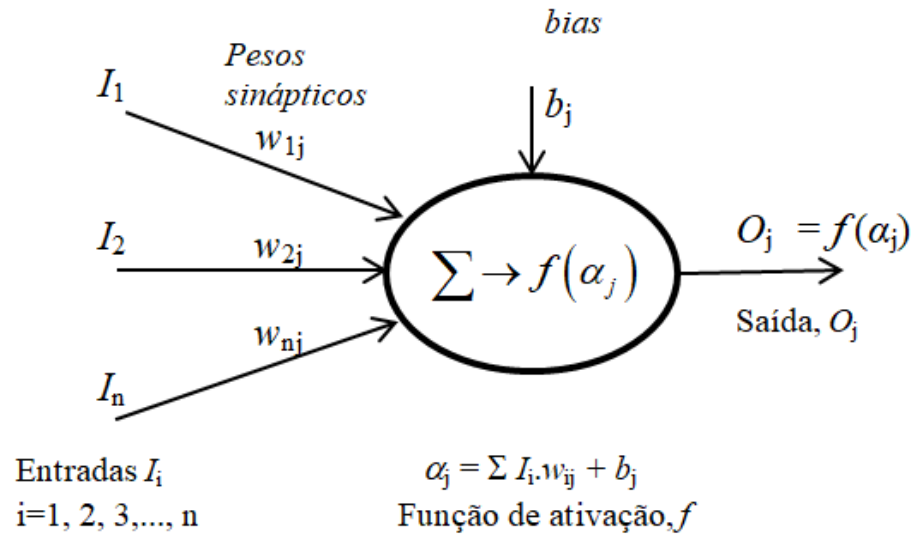
O processamento da informação em um neurônio biológico ocorre da seguinte forma: o sinal vindo de outro neurônio entra pelos dendritos e é transmitido para o corpo celular, local em que o processamento metabólico acontece. O axônio tem a função de transmitir o sinal nervoso do corpo celular para as extremidades da célula, os terminais do axônio, que se ligam nos dendritos do neurônio seguinte formando as sinapses.

Uma vez que os impulsos nervosos chegam ao corpo celular, ocorre a soma dos estímulos que pode, ou não, provocar a geração de um estímulo de saída do neurônio. Para que essa saída seja produzida, é necessário que o impulso gerado seja maior ou igual a um determinado limiar (Braga *et al.*, 2007). Por conter dezenas de bilhões de neurônios, e muitas conexões complexas entre eles, o cérebro humano se tornou capaz de aprender, e principalmente, a relacionar as respostas acerca de um determinado fenômeno com as variáveis que o influenciam.

3.3 O Neurônio Artificial

A estrutura de um neurônio artificial foi idealizada de modo que fosse similar à do neurônio biológico: ele possui entradas e saídas, sinapses, e um somador, como mostra a Figura 3.2.

Figura 3.2 – Estrutura de um neurônio artificial.



O funcionamento do neurônio artificial também simula o funcionamento do neurônio biológico. Assim, a informação, I_i, advinda do neurônio anterior, chega ao neurônio *j* através das sinapses. Essas ligações são caracterizadas por valores numéricos, os pesos sinápticos, w_{ij}, que ponderam a importância das entradas e são determinados na etapa de aprendizado da RNA. Além disso, cada neurônio é caracterizado por outro parâmetro, chamado *bias*, que confere versatilidade ao modelo e amplia o espaço de soluções, adequando a estrutura neuronal ao problema em questão (Braga *et al.*, 2007).

Os sinais que chegam ao neurônio *j*, depois de ponderados são somados e junto com o *bias*, formam o coeficiente de ativação, α_j, que é representado pela Eq. (2):

$$\alpha_j = \sum_{i=1}^{Z} I_i w_{ij} + b_j \qquad (2)$$

em que Z é o número de variáveis de entrada do neurônio.

3.3.1 Funções de Ativação

Como visto na seção anterior, cada neurônio artificial é caracterizado pela Eq. (2), que é calculado a partir das informações processadas anteriormente. No entanto, a maioria dos fenômenos naturais possui uma característica não linear que tal equação não leva em conta. Assim, as funções de ativação são necessárias para que essa não-linearidade seja levada em conta.

A saída do neurônio é então calculada por meio da aplicação da função da ativação sobre o valor do coeficiente de ativação do neurônio em questão, gerando a saída $O_j = f\left(\alpha_j\right)$. Tais funções podem assumir diversas formas, como tangente hiperbólica, secante hiperbólica, sigmoidal, etc. Normalmente, se utilizam funções contínuas, crescentes e limitadas por assíntotas horizontais para que a saída não atinja valores irreais, como infinito (Braga *et al.*, 2007).

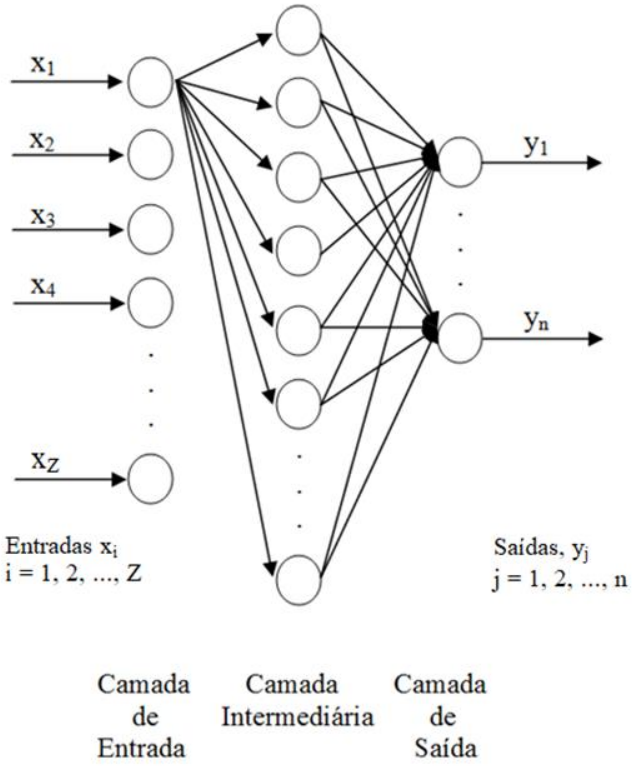
3.4 Estrutura e Equacionamento

As RNAs foram concebidas para atuar de forma similar ao cérebro humano. Assim, elas são um sistema matamático complexo, que consistem na distribuição de várias unidades de processamento simples, os neurônios artificiais, que têm a capacidade natural de acumular conhecimentos acerca de um fenômeno qualquer, a partir de exemplos e sem o prévio conhecimento das relações entre as variáveis que influenciam nesse processo (Haykin, 2005).

A unidade fundamental de uma RNA é o neurônio artificial, que por sua vez, se liga a outros neurônios formando um organismo computacional extremamente complexo em termos de funcionamento. A estrutura de uma RNA consiste em camadas de neurônios interligadas, através da qual as informações fornecidas são transmitidas unidirecionalmente. Uma RNA apresenta uma camada de entrada, uma de saída e, deve conter pelo menos uma camada intermediária (ou oculta).

A Figura 3.3 ilustra a estrutura de uma RNA com uma camada oculta. Para melhor visualização, somente a conexão do primeiro neurônio da camada de entrada com os da camada oculta foi apresentada. Porém, esse tipo de ligação se repete para todos os outros neurônios da camada de entrada.

Figura 3.3 – Estrutura geral de uma RNA com uma camada oculta.



Os neurônios podem se ligar camada a camada, como na Figura 3.3, ou então formar uma rede ainda mais complexa, se ligando a si próprios, como nas redes neurais recorrentes. No entanto, o tipo de aplicação é que determina qual a arquitetura (complexidade) de RNA é mais adequada (Braga *et al.*, 2007).

Como se pode notar da Figura 3.3, do ponto de vista prático, uma RNA nada mais é do que uma grande equação matemática, altamente não linear, e com diversos parâmetros ajustáveis, os pesos e *bias* de cada neurônio. Uma vez que esses parâmetros são ajustados, a equação para a variável de saída y_n de uma RNA com uma camada oculta pode ser escrita de uma forma simples, como apresentado na Eq. (3):

$$y_n = f\left(\sum_{j=1}^{W} w_{jk} f\left(\sum_{i=1}^{Z} w_{ij} x_i + b_j\right)_j + b_k\right)_n \tag{3}$$

em que: W é o número de neurônios da camada oculta, b_j e b_k são os *bias*, e w_{ij} e w_{jk} são os pesos sinápticos dos neurônios da camada oculta, e de saída, respectivamente. Para uma RNA com duas camadas ocultas, adiciona-se um termo a mais na Eq. (3):

$$y_n = f\left(\sum_{k=1}^{L} w_{kl} f\left(\sum_{j=1}^{W} w_{jk} f\left(\sum_{i=1}^{Z} w_{ij} x_i + b_j\right)_j + b_k\right)_k + b_l\right)_n \tag{4}$$

em que: W e L são o número de neurônios da primeira e da segunda camada oculta, respectivamente. b_l e w_{kl} são, respectivamente o *bias* e o peso sináptico dos neurônios da segunda camada oculta. Nesse caso, a primeira camada oculta corresponde à camada oculta única da RNA anterior.

Vale salientar que, em uma RNA, o número de neurônios das camadas de entrada e de saída é fixo e igual ao número de variáveis de entrada e saída, respectivamente. Além disso, não há aplicação de uma função de ativação para os neurônios da camada de entrada, visto que a função deles é apenas transmitir o valor das variáveis de entrada para os neurônios da camada seguinte, assim:

$$O_{j,\text{entrada}} = x_i \tag{5}$$

Em se tratando de um estudo envolvendo RNAs, em suma, há dois conjuntos de parâmetros que precisam ser ajustados. O primeiro diz respeito aos pesos e *bias* de cada neurônio e o segundo, ao número de camadas ocultas e ao número de neurônios em cada camada.

A obtenção do primeiro conjunto é relativamente fácil, visto que os parâmetros são ajustados automaticamente via a aplicação de um método numérico. Por outro lado, o grande desafio de se obter uma RNA reside no segundo conjunto, uma vez que a escolha desses parâmetros é uma tarefa que o operador da rede precisa executar. De antemão, não se sabe qual o tamanho ótimo da RNA para se descrever determinado fenômeno. Se a arquitetura for simples, o modelo pode não possuir a capacidade necessária, por outro lado, se

extremamente complexa, o ajuste dos parâmetros pode se tornar uma tarefa bastante onerosa visto que mais tempo computacional será necessário para ajustar os parâmetros (Braga *et al.*, 2007).

O método de obtenção da estrutura ótima que se deseja geralmente é o do tipo teste-e-erro. Desse modo, sugere-se um número de camadas ocultas e varia-se o número de neurônios nessa camada (por exemplo: de 2 a 30, considerando-se apenas os números pares). Em seguida, avalia-se o desempenho de todas as RNAs com relação à sua capacidade de predição. Assim, se o resultado obtido por alguma estrutura proposta for satisfatório, o estudo se encerra. Caso necessário, se acrescenta uma nova camada oculta e se propõem diferentes combinações de números de neurônios na primeira e segunda camada. Novamente, se avalia o desempenho de tais estruturas. O mesmo é feito caso haja a necessidade de uma terceira camada oculta, o que quase sempre não é necessário. Detalhes sobre a obtenção do modelo neural ótimo serão abordados nas Seções a seguir.

3.5 Treinamento das Redes Neurais Artificiais

A obtenção de uma RNA capaz de descrever um determinado fenômeno é realizada a partir da apresentação de exemplos; um conjunto de dados de fatores (variáveis) independentes (entrada) e dependentes (saída), em número suficiente tal que a RNA aprenda a relação que existe – o que lembra muito a forma de como o cérebro humano funciona.

Assim, a principal etapa do trabalho com RNAs é o aprendizado. Nessa etapa, os dados do problema a ser modelado são apresentados à RNA para que a rede possa aprender sobre ele. Nesse caso, o termo ‘aprender’ se refere ao ajuste matemático dos pesos sinápticos e dos *bias* da RNA. Ao fim do processo, os parâmetros ajustados representam o conhecimento que a RNA adquiriu.

O aprendizado de uma RNA é realizado por meio de um algoritmo de otimização, que ajusta os pesos sinápticos com o objetivo de se minimizar o erro entre os valores calculados pela rede e os experimentais. O ajuste dos pesos se dá da seguinte forma: sabendo que na iteração *i*, os pesos e *bias* sejam representados pelos vetores $\mathbf{w}(i)$ e $\mathbf{b}(i)$, o cálculo do valor dos pesos e *bias* da iteração seguinte é realizado segundo as equações:

$$\mathbf{w}(i+1) = \mathbf{w}(i) + \Delta\mathbf{w}(i) \tag{6}$$

$$\mathbf{b}(i+1) = \mathbf{b}(i) + \Delta\mathbf{b}(i) \tag{7}$$

Existem diversas formas de se calcular $\Delta\mathbf{w}(\mathrm{i})$ e $\Delta\mathbf{b}(\mathrm{i})$, dentre elas se encontram os algoritmos de aprendizado supervisionado e não supervisionado, sendo que a primeira é a forma mais comumente utilizada para se treinar uma RNA. No aprendizado supervisionado, a saída calculada pela RNA é continuamente comparada aos dados experimentais na fase de ajuste dos pesos (Haykin, 2005; Braga *et al.*, 2007).

3.5.1 Métodos de Otimização

Na etapa de treinamento de uma RNA, os pesos e *bias* são modificados até que um critério de parada seja satisfeito. Normalmente, tais valores são alterados de modo a se minimizar uma função objetivo que pode assumir diferentes formas (Haykin, 2005; Braga *et al.*, 2007). Sendo assim, o problema de ajuste dos parâmetros de uma RNA é um problema de otimização numérica.

Existem diversos algoritmos matemáticos desenvolvidos para esse fim e cada um possui suas vantagens e desvantagens. Neste trabalho, três métodos matemáticos foram usados no treinamento das RNAs: Levemberg-Marquardt, Powell e *Resilient Backpropagation*. Nas seções abaixo, os detalhes teóricos mais importantes de cada método serão abordados para a compreensão do leitor.

3.5.1.1 Método Back-Propagation

O *back-propagation* foi o primeiro método proposto para se treinar uma RNA. É um método do tipo gradiente descendente e é o mais utilizado, normalmente em conjunto com melhoramentos de outros autores. Nesse método, o treinamento da rede é dividido em duas etapas: na etapa *forward*, a saída da rede é calculada para um determinado conjunto de pesos sinápticos. Já, na etapa *backward*, as saídas calculadas e fornecidas são utilizadas para atualizar os pesos sinápticos para a próxima iteração. Nesse algoritmo a atualização dos

parâmetros (seja peso ou *bias*), *t*, da RNA em uma iteração *i* é realizada de acordo com a equação:

$$\Delta t(i) = -\nabla E(t(i)) \tag{8}$$

A dedução e mais detalhes matemáticos sobre esse método podem ser encontrados em Braga *et al.* (2007).

3.5.1.2 Método Levemberg-Marquardt

O método Levemberg-Marquardt (Marquardt, 1963) é baseado no método de Newton e é um dos mais utilizados para se treinar RNAs. É utilizado juntamente com o *back-propagation*, devido à sua maior eficiência (Braga *et al.*, 2007). A diferença em relação ao método *Backpropagation* é que a atualização dos pesos sinápticos é realizada utilizando a equação:

$$\mathbf{\Delta t}(i) = -\left[\nabla^2 E(t(i)) + \mu \mathbf{I}\right]^{-1} \nabla E(t(i)) \tag{9}$$

em que: $\nabla^2 E(t(i))$ e $\nabla E(t(i))$ são a matriz Hessiana e o gradiente do erro em relação ao parâmetro *t* na iteração *i*, μ é a taxa de aprendizado, e **I** é a matriz identidade. A taxa de aprendizado é outro parâmetro que influencia no resultado final do treinamento. Neste trabalho, ela não foi avaliada e seu valor foi definido de acordo com a configuração padrão do *software* utilizado.

3.5.1.3 Método Resilient Back-Propagation

O método *Resilient Back-Propagation* foi proposto por Riedmiller e Braun (1992a) e seu objetivo era contornar um problema recorrente, especialmente quando se trabalha com variáveis de saída normalizadas no intervalo de [-1;1]. Em certos casos, quando a saída desejada é 1 (ou 0) e a saída calculada pela RNA for 0 (ou 1), a derivada do erro em relação ao parâmetro pode ser muito próxima de 0. Isso pode ser um problema, pois se a derivada tiver uma magnitude baixa, o valor da atualização do parâmetro também será.

Nesse método, que é aplicado em conjunto com o método *Backpropagation* a atualização dos parâmetros da RNA é realizada considerando-se, apenas, o sinal da derivada e o valor da atualização é definido através das Eqs. (10) e (11) (Braga *et al.*, 2007):

$$\Delta t(i) = \begin{cases} -\Delta(i); & \text{se} \quad \dfrac{\partial E(i)}{\partial t} > 0 \\ +\Delta(i); & \text{se} \quad \dfrac{\partial E(i)}{\partial t} < 0 \\ 0; & \text{se} \quad \dfrac{\partial E(i)}{\partial t} = 0 \end{cases} \tag{10}$$

$$\Delta(i) = \begin{cases} \eta^{+}\Delta(i-1); & \text{se} \quad \dfrac{\partial E(i-1)}{\partial t}\dfrac{\partial E(i)}{\partial t} > 0 \\ \eta^{-}\Delta(i-1); & \text{se} \quad \dfrac{\partial E(i-1)}{\partial t}\dfrac{\partial E(i)}{\partial t} < 0 \\ \Delta(i-1); & \text{se} \quad \dfrac{\partial E(i)}{\partial t} = 0 \end{cases} \tag{11}$$

Os valores de η^{-} e η^{+} foram determinados empiricamente e valem 0,5 e 1,2, respectivamente (Riedmiller; Braun, 1992b).

3.5.1.4 Método de Powell-Beale

Diferentemente dos métodos descritos nas seções 3.5.1.1, 3.5.1.2 e 3.5.1.3, o método de Powell-Beale (Beale, 1972; Powell, 1977) encontra o mínimo de uma função sem a necessidade de se calcular derivadas de segunda ordem. Isso é computacionalmente vantajoso pelo fato de que é necessário se armazenar poucas matrizes ao longo da otimização dos parâmetros. Esse método se baseia no fato de que, no ponto mínimo de uma direção **u**, o gradiente da função *f* é perpendicular, ou conjugado, a essa direção.

A atualização dos parâmetros da RNA segundo o algoritmo de Powell (1977) se dá de acordo com a Eq. (12):

$$\Delta t(i) = \alpha_i d_i \tag{12}$$

em que: d_i é uma equação que depende do gradiente do erro e α_i é uma constante. Essas duas variáveis serão abordadas nos parágrafos subsequentes.

Powell (1977) partiu do pressuposto de que a cada número determinado de iterações é necessário reiniciar a direção de procura para o negativo do gradiente do erro. Nesse caso, a reinicialização ocorre a cada (número de pesos + número de *bias*) iterações. No algoritmo proposto, o autor recomenda se analisar a validade da inequação:

$$\left| g_{i-1}^{T} g_i \right| > 0,2 \left\| g_i \right\|^2 \tag{13}$$

em que: $g_i = \dfrac{\nabla E(i)}{\partial t}$. Quando a Eq. (13) é verdadeira, isso significa que a diferença entre os gradientes da iteração $i - 1$ e i praticamente estão na mesma direção. Assim, o algoritmo reinicia a direção de busca, considerando que d_i é o negativo do gradiente (Powell, 1977).

A grande vantagem desse método é que ele minimiza cada direção de modo independente. Assim, uma vez que a função tenha sido minimizada para um dado conjunto de direções conjugadas, não é necessário se realizar uma nova minimização nessas direções. Maiores detalhes sobre o método pode ser encontrado em Powell (1977).

3.6 Pós-Treinamento

Após a etapa de treinamento das RNAs é necessário verificar se a equação obtida possui a capacidade de prever novos casos com certo nível de exatidão. Usualmente, antes de um modelo neural ser considerado apto para ser utilizado é realizada duas etapas de verificação: a validação e o teste. Nessas duas etapas são utilizados como dados de entrada, um conjunto de pontos experimentais não apresentados à RNA na etapa de treinamento (Haykin, 2005; Braga *et al.*, 2007).

As duas etapas se diferem em um único ponto: a etapa de validação também é utilizada como critério de parada no treinamento de uma RNA, já a etapa de teste é realizada após a convergência ter ocorrido. Isso é realizado para evitar o que se define como sobreajuste, que acontece quando a RNA fica viciada apenas nos valores apresentados na etapa de treinamento. Uma consequência disso é que a equação só consegue prever

adequadamente os casos em que os valores das variáveis de entrada sejam muito próximos aos apresentados na etapa de treinamento (Braga *et al.*, 2007).

Como mencionado anteriormente, os dados experimentais coletados são divididos em três subconjuntos: treinamento, validação e teste. Assim, para se evitar o sobreajuste recomenda-se acompanhar o comportamento dos dados marcados como validação ao longo do treinamento da RNA. Uma vez que a função objetivo da validação atinge um ponto de mínimo e passa a aumentar, recomenda-se parar o treinamento após um número de iterações definido pelo usuário e considerar, como resultado final, a iteração cuja validação apresentou o ponto de mínimo da sua função objetivo.

Por fim, executa-se a predição da RNA com os dados identificados como teste. Caso essa etapa apresente resultados em níveis adequados, considera-se que o modelo neural está apto para ser aplicado no cotidiano.

3.7 Aplicações como Sistema de Suporte à Decisão Clínica

A simulação é uma ferramenta que pode ser utilizada por médicos para se tomar decisões precisas e efetivas a tempo de se salvar vidas. O uso de RNAs para prever enfermidades, diagnosticar, ou classificar pacientes, tem se mostrado bastante promissor, e diversos estudos envolvendo essa ferramenta podem ser encontrados na literatura.

A presença de coágulos de sangue em artérias foi modelada via RNAs por Lela *et al.* (2014). Nesse estudo, os autores consideraram 10 variáveis de entrada, sendo uma combinação de sintomas e fatores de risco. Os resultados mostraram que uma RNA com uma camada oculta contendo 20 neurônios são promissores, apresentando alto nível de acurácia.

Latterie e Borseth (2014) aplicaram RNAs para predizer a dose de gonadotrofina, gonadotrofina coriônica humana e o cancelamento do ciclo, durante a estimulação ovariana na fertilização *in vitro*. Como variáveis de entrada, os autores utilizaram 15 variáveis clínicas independentes. Os resultados obtidos pelos autores mostraram que as RNAs são capazes de predizer com confiança as três variáveis e se mostraram de acordo com as decisões tomadas pela equipe de médicos.

Kumar *et al.* (2014) utilizaram RNAs combinadas com o algoritmo colônia de formigas no diagnóstico de pacientes com diabetes. Nesse estudo, os autores consideraram um conjunto com oito variáveis de entrada e outro, obtido através de um algoritmo de seleção de dados, que continha quatro variáveis de entrada. Ambos os conjuntos consideravam dados físicos e clínicos do paciente. O resultado obtido pelos autores mostrou que, para o segundo conjunto, uma RNA com quatro neurônios na camada oculta era capaz de prever os casos de diabetes com 90 % de precisão.

Kojuri *et al.* (2015) compararam o desempenho entre RNAs do tipo *back-propagation* e do tipo função de base radial para prever casos de infarto do miocárdio. Nesse estudo, os autores consideraram como variáveis de entrada o histórico clínico, exame físico, laboratorial, e de eletrocardiogramas dos pacientes. Os resultados mostraram que RNAs do tipo *back-propagation* apresentaram melhor desempenho (97 %) em relação às de função de base radial (90 %).

Süt e Çelik (2012) aplicaram RNAs para predizer a mortalidade de pacientes que tiveram infarto do miocárdio. Em seu estudo, os autores compararam o desempenho de seis algoritmos de treinamento diferentes, considerando oito variáveis de entrada. Os resultados mostraram que a predição depende do algoritmo de treinamento utilizado, sendo que a precisão variou entre 60 % e 81 %.

Abedi *et al.* (2017) desenvolveram uma RNA para reconhecer e diagnosticar isquemia cerebral aguda e diferencia-la de um infarto.Os autores consideraram 19 fatores clínicos como variáveis de entrada e os resultados mostraram que a RNA possui capacidade de acerto de 92 % para o reconhecimento de isquemia cerebral aguda.

Saha e Mandal (2017) usaram RNAs do tipo perceptron na predição de casos de Dengue em pacientes indianos, considerando os sintomas declarados pelos pacientes como variáveis de entrada. Os autores mostraram que a RNA é efetiva na detecção de casos de dengue.

Badnjević *et al.* (2016) treinaram RNAs visando classificar pacientes que possuem asma a partir de exames clínicos, como variáveis de entrada. Os autores treinaram

diversas estruturas, com dados de 1800 pacientes. Os resultados obtidos mostraram que as RNAs conseguem classificar os pacientes asmáticos, com precisão de 97 %.

Apesar de muito utilizadas na medicina, estudos envolvendo casos de trombose ainda são muito incipientes. São poucos os trabalhos encontrados sobre tal assunto. A predição de casos de TVP e EP usando RNAs foi realizada por Ghavami e Kapur (2011), considerando 20 variáveis de entrada, dentre elas: duração da internação, peso, IMC, variáveis hematológicas, etc. Os resultados desse trabalho mostraram que as RNAs conseguem prever tais casos com erros abaixo de 10 %.

Mais recentemente, Fei *et al.* (2017b) aplicaram RNAs para prever a incidência de trombose venosa mesentérica em pacientes com pancreatite aguda. Os autores treinaram uma RNA *feedforward backpropagation* com dados de 72 pacientes, e 11 variáveis de entrada (dados do paciente e fatores clínicos). Os resultados obtidos pelos autores mostraram que uma RNA com 9 camadas intermediárias era melhor do que o modelo de regressão logística empregado usualmente. Em outro trabalho, Fei *et al.* (2017a) treinaram redes neurais do tipo base radial visando predizer o risco de trombose de veia porta em pacientes com pancreatite aguda e obtiveram resultados similares.

Por fim, Qatawneh *et al.* (2017) treinaram diversas RNAs com até 5 camadas ocultas para predizer o risco de trombose em pacientes hospitalizados. Como variáveis de entrada, os autores utilizaram diversas características do pacientes, da doença responsável pela internação, e fatores genéticos. Os resultados obtidos confirmaram que as RNAs podem ser uma alternativa para esse fim, uma vez que a precisão foi de 81 %.

Outros estudos sobre a aplicação das RNAs para diagnóstico e prevenção de diversas doenças também podem ser encontrados nos trabalhos de revisão de Kutamari e Sunita (2013), e de Parveen *et al.* (2016). Até a finalização deste trabalho nenhum grupo de pesquisa no Brasil reportou a utilização de RNAs visando o diagnóstico ou prognóstico, relacionado a qualquer enfermidade.

3.8 Conclusões

Este Capítulo mostrou que as RNAs podem ser uma alternativa viável para a modelagem de fenômenos complexos. Uma vez que não é necessário se conhecer como as relações entre as variáveis dependentes e independentes se dão, a obtenção de um modelo neural e sua utilização pode ser bastante simplificada. Para o caso da predição da TR, a obtenção de um modelo neural pode ser bastante interessante, visto que se conhecer a relação matemática fenomenológica entre as variáveis é extremamente complicado e trabalhoso.

4 ANÁLISE DE COMPONENTES PRINCIPAIS

Neste trabalho, a Análise de Componentes Principais foi utilizada visando identificar quais são os fatores preditivos da TR. Por esse motivo, nesse Capítulo serão abordados os seus princípios básicos e serão apresentadas diferentes aplicações da técnica na área da saúde.

4.1 Introdução

A identificação dos fatores mais importantes que influenciam um determinado fenômeno pode ser realizada de três maneiras: a partir da modelagem fenomenológica rigorosa do problema, da análise do índice de correlação, ou de análise estatística multivariada. O primeiro método é o mais rigoroso, uma vez que é realizado utilizando-se uma equação desenvolvida especialmente para o fenômeno em questão, levando em consideração todas as suas propriedades. Porém, é necessário profundo conhecimento do fenômeno e, muitas vezes, de matemática avançada para ser realizado. O segundo possui a vantagem da possibilidade de ser realizado de modo empírico. No entanto, essa pode ser uma tarefa onerosa e custosa dependendo da resposta que se deseja, pois são necessários diversos experimentos (muitas vezes fatoriais). O terceiro método se torna uma alternativa interessante, pois consiste em se analisar estatisticamente a distribuição de um conjunto de dados obtido aleatoriamente.

A Análise de Componentes Principais (ACP) é uma técnica de análise estatística multivariada que tem dois objetivos: i) identificar padrões e fatores importantes em um determinado fenômeno e, ii) reduzir a dimensionalidade de um conjuntos de dados sem perda significativa de informação. Como resultado final, a ACP gera as chamadas componentes principais (CPs) que são combinações lineares dos valores originais das variáveis. Esse método foi desenvolvido matematicamente por Hotelling (1933a, b) e a partir de então utilizado em diversas áreas da ciência na análise e pré-tratamento de dados.

A base matemática da ACP se deve ao fato de que o vetor que maximiza a forma quadrática associada a uma matriz simétrica é a matriz dos autovetores ordenados de tal forma que seus autovalores correspondentes, e consequentemente a variância de cada componente,

se apresentam em ordem decrescente. Geometricamente, as CPs consistem em uma rotação de eixos de tal forma a selecionar novos eixos coordenados que representem a direção de máxima variabilidade (Hotelling, 1933a, b; Härdle; Simar, 2007).

4.2 Dedução Matemática

Considerando-se uma matriz de dados **X**, que contém *n* número de pontos experimentais ou dados coletados (linhas), e *m* número de fatores (colunas), é possível obter uma matriz de Componentes Principais, **CP**, que possui como característica principal a correlação zero entre suas componentes *j*. O cálculo das CPs se é realizado seguindo os passos descritos nos parágrafos a seguir.

Primeiramente, os valores originais das variáveis precisam ser padronizados centrando a média dos valores em zero, com variância igual a um, eliminando assim, a heterogeneidade dos dados especialmente no tocante à ordem de grandeza. Para isso, se utiliza a equação:

$$Z_{i,j} = \sigma_j^{-1/2}(X_{i,j} - \mu_j); i = 1,\ n;\ j = i, m \tag{14}$$

em que: *m* é o número total de fatores, $X_{i,j}$ é o *i-ésimo* dado experimental da variável *j*, $Z_{i,j}$ é o valor padronizado do dado $X_{i,j}$ e, μ_j e σ_j são a média e a variância da variável *j*, respectivamente.

Assim, a partir da matriz *Z*, se pode calcular a matriz variância-covariância, $\mathbf{s}$, dos fatores utilizando-se a equação:

$$s_{Z_j Z_k} = \frac{\sum_{j,k=1}^{n} Z_j Z_k - \sum_{j=1}^{n} Z_j \sum_{k=1}^{n} Z_k}{n} \tag{15}$$

em que: *j* e *k* são fatores e *n* é o número total de dados experimentais. A matriz *s* é uma matriz quadrada em que a diagonal principal corresponde à variância do fator Z_j e os demais valores são a covariância entre dois fatores Z_j e Z_k.

Uma vez calculada a matriz *s*, pode-se calcular seus autovetores e autovalores correspondentes, resolvendo-se as equações:

$$\left|\mathbf{s} - \lambda\mathbf{I}\right| = 0 \tag{16}$$

$$(\mathbf{s} - \lambda\mathbf{I})\mathbf{a} = 0 \tag{17}$$

em que: **I** é a matriz identidade de ordem *n*, λ são os autovalores e **a** é a matriz de autovetores da matriz **s**. Da Eq. (17), pode-se notar que cada autovetor a_i se relaciona diretamente com um autovalor λ.

Para se calcular as CPs, uma vez obtidos os valores de λ e **a**, é necessário ordenar os autovalores em ordem decrescente, assim como reorganizar a matriz **a**, de modo que os autovetores mantenham a correspondência com os autovalores. Por fim, as CPs são geradas pela multiplicação da matriz dos dados padronizados com a dos autovetores:

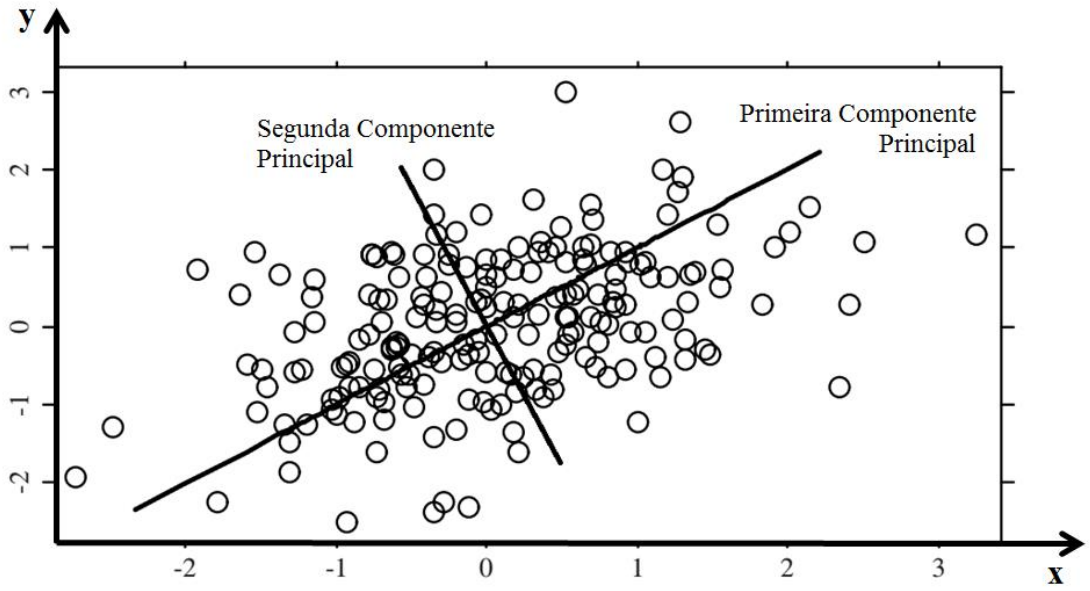
$$\mathbf{CP} = \mathbf{Z} \cdot \mathbf{a} \tag{18}$$

Da Eq. (18), nota-se que as CPs são combinações lineares das variáveis originais ponderadas pelos coeficientes dos autovetores e é uma matriz *n* x *m*. A Figura 4.1 apresenta uma representação gráfica da transformação de eixos da ACP.

O princípio básico da ACP é ilustrado nesssa Figura, onde se pode notar que a primeira CP corresponde à direção de maior variância dos dados. A segunda CP é ortogonal em relação à primeira e corresponde à segunda direção com a maior variância dos dados. Essa análise segue de modo análogo para todas as outras CPs.

Uma grande vantagem da ACP é a possibilidade de redução de dimensionalidade dos dados, ou seja, nem todas as CPs calculadas são necessárias para representar o fenômeno em questão. Na prática, basta utilizar as CPs cujos autovalores são superiores que a unidade. Todas as outras CPs podem ser descartadas, pois um autovalor menor que um indica que a importância é menor do que as próprias variáveis originais.

Figura 4.1 - Representação gráfica das CPs.



Fonte: Adaptado de Härdle e Simar (2007).

Outra forma de se reduzir a dimensionalidade de um conjunto de dados é avaliar quais fatores são importantes naquelas CPs com autovalor superior a um. Uma vez obtidos **a** e λ, pode-se efetuar a análise. Isso é realizado comparando-se o valor do coeficiente de correlação entre as variáveis originais e as CPs, $\rho_{X_i Z_j}$,dado pela Eq. (19) (Härdle; Simar, 2007). Considera-se que um fator é importante em uma determinada CP quando $\left|\rho_{X_i Z_j}\right| > 0{,}5$. Uma consequência direta é que tal fator poderá ser considerado importante para o fenômeno.

$$\rho_{X_i Z_j} = a_{ij}\sqrt{\frac{\lambda_i}{s_{X_i X_j}}};\ \ i,\ j = 1,\ 2,\ldots,m \tag{19}$$

É importante salientar que é necessário se ter um conjunto completo de dados para se utilizar a ACP. Caso isso não seja verdadeiro, é necessário se empregar algum método para inserção de dados em lacunas de uma tabela. Na Seção 6.2, esse tópico será abordado em maiores detalhes.

4.3 Aplicações na Identificação de Fatores Preditivos de Doenças

A ACP tem sido raramente empregada na classificação de padrões na área da saúde e poucos trabalhos recentes podem ser encontrados na literatura:

Kutcher *et al.* (2013) aplicaram ACP para identificar padrões de anormalidade na coagulação de pacientes com coagulopatia traumática aguda. Os autores acompanharam 163 pacientes e os resultados mostraram que 3 de 10 CPs bastam para descrever tal fenômeno. Nestas CPs, apenas 5 dos 10 fatores selecionados são importantes.

Chin *et al.* (2014) empregaram ACP para determinar grupos de fatores relationados à coagulopatia induzida por trauma. Em seu estudo, os autores acompanharam 98 pacientes em um período de três anos e os resultados mostraram que a técnica foi capaz de identificar três grupos principais de fatores: coagulopatia global com queda de plaquetas e fibrinogênio, ativação da proteína C e, hiperfibrinólise. Juntas essas três CPs explicaram mais do que 93 % da variância dos dados.

Thorpe *et al.* (2016) investigaram a dieta da população australiana entre 55-65 anos, considerando 52 grupos de alimentos e 3959 pacientes, usando ACP. Os resultados obtidos motraram que para os homens 4 CPs eram suficientes para descrever os padrões de dieta e que para as mulheres, 2 CPs. Os autores também foram capazes de identificar quais eram os grupos principalmente consumidos pelos pacientes.

Vavougios *et al.* (2016) investigaram os padrões das comorbidades da apnéia via ACP. Nesse trabalho, os resultados mostraram que os pacientes podem ser separados em 6 diferentes grupos: saudável, médio, moderado sem comorbidades, moderado com comorbidades, severo sem comorbidades e, severo com comorbidades severas. Ainda, os autores identificaram os principais fatores que afetam cada grupo.

4.4 Conclusões

Este Capítulo mostrou que técnicas estatísticas multivariadas podem ser grandes aliadas na redução da dimensionalidade de dados bem como na identificação de variáveis importantes para um determinado fenômeno. A ACP pode ser uma ferramenta poderosa na

formulação e simplificação de modelos matemáticos, especialmente em aplicações na Medicina, que geralmente possuem diversos fatores como variáveis independentes e que apresentam relações extremamente complexas.

5 PROCEDIMENTO PARA OBTENÇÃO DO MODELO NEURAL

5.1 Coleta das Informações

Todos os dados utilizados neste trabalho foram coletados no banco de dados da área de Trombose do Setor de Hemostasia do Hemocentro da Unicamp. Ao todo, foram obtidas informações sobre 307 pacientes com um primeiro antecedente de trombose entre janeiro de 2009 e agosto de 2016. Foram considerados como elegíveis os pacientes maiores de 18 anos com trombose venosa de membros inferiores ou no sistema nervoso central, e assim 72 pacientes foram excluídos por não satisfazer essas condições, restando 235. Conforme exposto no Capítulo 2, diversas variáveis podem influenciar na TR. Assim, foram considerados como variáveis independentes os fatores apresentados na Tabela 5.1.

Tabela 5.1 – Fatores considerados neste estudo.

Sexo(Masculino/Feminino)	Mutação G20210A no gene da protrombina (Heterozigoto/Não)	Hemácias (μL^{-1})	Glicose ($mg.dL^{-1}$)
Idade (anos)	Factor VIII ($IU.dL^{-1}$)	Hemoglobina ($mg.dL^{-1}$)	Creatinina ($mg.dL^{-1}$)
Embolia Pulmonar (Sim/Não)	Atividade de PC ($mg.dL^{-1}$)	Hematócrito (%)	Proteína C reativa ($mg.dL^{-1}$)
Lado da trombose no membro inferior (Direito/Esquerdo)	Atividade de PS ($mg.dL^{-1}$)	Distribuição de tamanho dos eritrócitos (%)	Hipertensão arterial (Sim/Não)
Localização da trombose no membro inferior (Distal/Proximal)	Atividade de AT ($mg.dL^{-1}$)	Plaquetas (μL^{-1})	Diabetes (Sim/Não)
Provocada ou espontânea	Síndrome do anticorpo antifosfolipídeo (Sim/Não)	Volume médio das plaquetas (fL)	Dislipidemia (Sim/Não)
Tempo de anticoagulação (meses)	Índice de massa corporal (IMC)	Colesterol total ($mg.dL^{-1}$)	Insuficiência Renal (Sim/Não)
Uso de anticoagulante (Sim/Não)	Tabagismo (Sim/Não)	Lipoproteína de alta densidade (HDL) ($mg.dL^{-1}$)	Cancer (Sim/Não)
D-dímero ($ng.mL^{-1}$)	Terapia hormonal (para mulher) (Sim/Não)	Lipoproteína de baixa densidade (LDL) ($mg.dL^{-1}$)	Trombo residual (Sim/Não)
FV Leiden (Heterozigoto/Não)	Leucócitos (μL^{-1})	Triglicérides ($mg.dL^{1}$)	

Os fatores considerados neste trabalho foram obtidos antes do episódio de retrombose e se dividem em dois grupos: i) características do primeiro episódio trombótico e do seu tratamento e, ii) dados clínicos do paciente, totalizando 39 fatores. Para a coleta de dados, foram utilizados dois bancos de dados diferentes. No "Sistema de Hematologia do Hemocentro da Unicamp" foram coletadas informações sobre o paciente, tais como: idade, sexo, histórico da primeira e segunda trombose, e informações sobre trombo residual, localização da trombose no membro inferior – obtidas através de exames de ultrassom, e resultados de exames laboratoriais de trombofilia e hemograma. Já, no denominado "Sistema do Hospital de Clínicas", foram coletadas as informações referentes aos exames laboratoriais metabólicos. Considerou-se que a trombose era provocada naqueles casos que apresentaram um episódio trombótico na presença de qualquer fator de risco: gestação, puerpério, anticoncepcional hormonal, repouso prolongado, câncer, cirurgia, trauma local, viagem aérea longa. Acredita-se que esse conjunto de variáveis seja suficiente para estabelecer uma relação matemática que classifique quais pacientes apresentarão TR. Ademais, vale salientar que os dados laboratoriais são de fácil obtenção, e que a maior parte deles são de rotina na prática clínica, e determinados por exame de sangue.

Para os passos seguintes, a matriz de dados foi adaptada de modo a conter apenas valores numéricos. Primeiro, as variáveis cujos valores eram na forma "Sim/Não" foram alteradas para a forma "+1/-1", respectivamente. Outras variáveis não foram apresentadas nessa forma, como por exemplo, o sexo do paciente cujo valor é "homem/mulher". Para esse tipo de variável, cada opção foi codificada com um número. Na Tabela 5.2 é apresentada a codificação utilizada para todas as variáveis não-numéricas. A variável resposta deste trabalho é a retrombose do paciente. Ela possui a forma "Sim/Não" e foi codificada com os valores "+1/-1", respectivamente.

5.2 Análise de Componentes Principais

Uma vez que o número de fatores selecionados é relativamente alto, a técnica de ACP foi utitilizada com o objetivo de determinar quais são os fatores predominantes para a população em questão. Esse passo é importante para facilitar a formulação de um modelo matemático mais simples.

Para isso, sobre a matriz dos dados coletados foram aplicadas as Eqs. (14)-(19), de modo que as CPs foram determinadas, assim como a correlação entre os fatores e cada CP obtida. A partir dos resultados, os principais fatores que influenciam a TR puderam ser determinados, considerando-se uma correlação acima de 0,5.

Tabela 5.2 – Codificação das variáveis de entrada não-numéricas.

Fator	**Entrada da RNA**
Sexo	Masculino = 1; Feminino = -1
Embolia pulmonar	Sim = 1; Não = -1
Lado da trombose no membro inferior	Lado direito = 1 Lado esquerdo = -1
Localização da trombose no membro inferior	Proximal = 1; Distal = -1
Provocada ou espontânea	Provocada = 1; Espontânea = -1
Uso de anticoagulante	Sim = 1; Não = -1
FV Leiden	Negativo = 1; Heterozigoto = -1
Mutação G20210A no gene prototrombina	Negativo = 1; Heterozigoto = -1
Síndrome do anticorpo fosfolipídeo	Negativo = 1; Positivo = -1
Tabagismo, terapia hormonal (para mulheres), hipertensão arterial, diabetes, dislipidemia, câncer, trombo residual	Sim = 1; Não = -1

5.3 Ajuste dos Modelos Neurais

O principal objetivo deste trabalho é obter modelos matemáticos neurais visando predizer quais pacientes apresentarão TR e quais não. A Figura 5.1 ilustra o procedimento adotado neste trabalho, em que foram considerados diferentes conjuntos de variáveis de entrada, gerando quatro modelos:

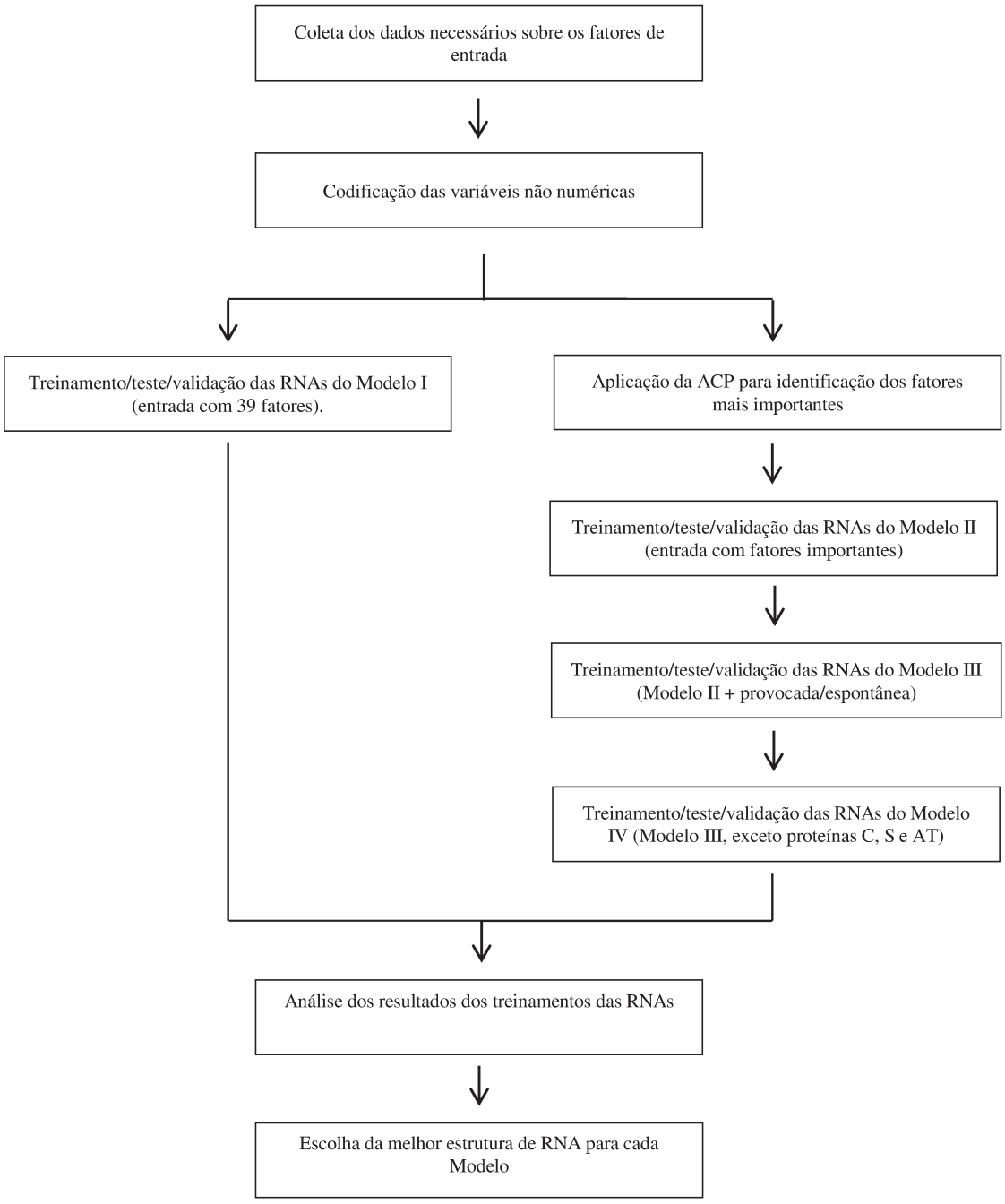
Modelo I: Os 39 fatores coletados;

Modelo II: Os principais fatores determinados através da ACP;

Modelo III: As variáveis do Modelo II, incluindo-se a variável: trombose provocada ou espontânea;

Modelo IV: As variáveis do Modelo III, excluindo-se as variáveis: proteína C, proteína S e Antitrombina.

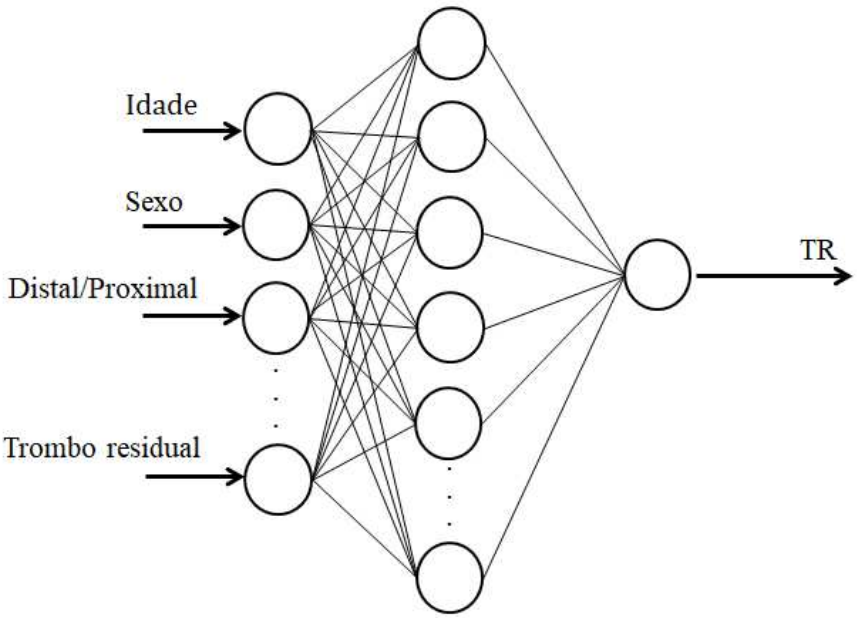
Figura 5.1 – Diagrama do procedimento para a obtenção dos modelos neurais.



Os Modelos III e IV surgiram como um desdobramento do Modelo II e os motivos dessa escolha serão abordados na Seção 6.4 e em detalhes no Capítulo 9. Ainda, vale salientar que as RNAs foram treinadas com os valores originais dos fatores, e não com as componentes principais calculadas usando a ACP. Essa forma de inserção das variáveis de entrada das RNAs visa facilitar a sua posterior utilização pelo usuário.

Diversas estruturas de RNAs com 1, 2 e 3 camadas ocultas foram testadas e diferentes algoritmos de otimização foram considerados (Levemberg-Marquardt, Powell-Beale e *Resilient Backpropagation*). Para todos os Modelos foram utilizados a mesma função de ativação em todos os neurônios (tangente hiperbólica), se minimizou a mesma função objetivo e se utilizou 70 % dos dados para treinamento, 15 % para validação e 15 % para teste. Esse amplo espaço de procura visava comparar o desempenho obtido nas diferentes situações. O critério de escolha do melhor modelo considerou o valor da função objetivo, bem como dos coeficientes de correlação obtidos nas etapas de treinamento, validação e teste. A Figura 5.2 ilustra a estrutura geral de uma RNA com uma camada oculta utilizada neste trabalho. Nessa Figura, cada neurônio da camada de entrada corresponde à informação de uma variável de entrada e o neurônio da camada de saída gera a resposta desejada.

Figura 5.2 – RNA com uma camada oculta utilizada neste trabalho.



6 PRINCIPAIS RESULTADOS DESTE TRABALHO

Os Capítulos 7 e 8 serão apresentados no formato do manuscrito submetido para publicação. No Capítulo 9, serão apresentados os resultados referentes aos dois últimos casos testados. Por esse motivo, para facilitar a leitura, nessa Seção será apresentado um resumo dos principais resultados encontrados.

6.1 Coleta de Dados e Análise da População

A Tabela 6.1 mostra as principais características dos eventos trombóticos observados nos dados coletados.

Tabela 6.1 – Principais características dos eventos trombóticos.

Intervalo entre a trombose e a recorrência	23.75 ± 19.12 meses	
Intervalo entre o fim do tratamento e a recorrência	15.87 ± 14.89 meses	
	Primeiro TEV	Segundo TEV
Homem	75	20
Mulher	161	29
Local	Número de pacientes	
Perna esquerda	108	23
Perna direita	71	15
Embolismo pulmonar (exclusivo)	42	9
Embolismo pulmonar	85	11
Sistema Nervoso Central	5	1
Ambas as pernas	6	1

Dos 235 pacientes cujos dados foram coletados, 49 apresentaram TR. Isso corresponde à 20,8 % da população amostrada. O primeiro episódio foi predominantemente na perna esquerda (45,7 %). Trinta e seis porcento apresentaram embolia pulmonar, seja exclusivamente ou associada a trombose de sítio, e pacientes com TVC foram 2,1 %. Quando se analisa a TR, observa-se que as porcentagens foram semelhantes: 46,0 %, 30,0 %, e 2,0 % para pacientes com TR na perna esquerda, direita e em bilateral, respectivamente. Vinte e dois

porcento apresentaram embolia pulmonar e 2,0 % TR no sistema nervoso central. Os dados mostram que 61 % dos pacientes apresentaram uma primeira trombose espontânea, enquanto que em 39 % a trombose foi provocada. Ainda, 40 % dos indivíduos que tiveram uma primeira trombose provocada apresentaram uma segunda TR (o que corresponde a 16 % do total de pacientes). Esse é um resultado importante, visto que se esperava que pacientes com trombose provocada não apresentassem um alto índice de recorrência.

Por fim, 13 pacientes da amostra tinham câncer. Desses, 3 apresentaram uma primeira trombose espontânea e apenas 1 teve retrombose. Tal indivíduo havia apresentado uma primeira trombose espontânea. No Capítulo 7, uma análise mais profunda dos demais resultados será apresentada.

6.2 Primeiro Manuscrito

O manuscrito apresentado no Capítulo 7 é intitulado: “Principal component analysis on recurrent venous thrombosis” e se refere ao estudo dos fatores preditivos da TR, realizado via aplicação da ACP. Dentro do contexto apresentado no Capítulo 4, os principais objetivos dessa etapa do trabalho foram:

1) Determinar o número de Componentes Principais suficientes para substituir o conjunto de dados original;
2) Determinar quais fatores possuem forte correlação em cada Componente Principal.

Todos os dados coletados foram utilizados nessa etapa, e assim foram consideradas as informações sobre os 39 fatores. Posterior à aplicação do método, os resultados mostraram que das 39 CPs, 18 são suficientes para descrever a TR. Analisando-se o coeficiente de correlação entre as variáveis e cada CP, determinou-se que 18 são os fatores que mais contribuem para a predição da TR. Tais fatores estão apresentados na tabela 5 do manuscrito (Table 5). Por fim, esta etapa do trabalho confirmou que a ACP é uma técnica poderosa na determinação de fatores preditivos de doenças ou suas complicações. No caso deste estudo, como ferramenta para predizer a recorrência de trombose venosa.

6.3 Segundo Manuscrito

O manuscrito apresentado no Capítulo 8 é intitulado: "New clinical decision support system for recurrent venous thrombosis" e se refere à obtenção de novos sistemas de suporte à decisão clínica, baseado em RNAs, para predição de TR. Assim, dentro do contexto apresentado nos Capítulos 3 e 5, os principais objetivos dessa etapa do trabalho foram:

3) Comparar o desempenho das RNAs considerando dois conjuntos de variáveis de entradas:
   Modelo I: os fatores originais (39 variáveis);
   Modelo II: os principais fatores, considerando o resultado obtido via ACP (18 variáveis).
4) Treinar diferentes estruturas de RNAs com 1, 2 e 3 camadas;
5) Comparar o desempenho de diferentes algoritmos de otimização;

Para o Modelo I, os resultados mostraram que a escolha do algoritmo de otimização é crucial para o desempenho do treinamento das RNAs, uma vez que a precisão de uma mesma estrutura variou significativamente com o método utilizado. As RNAs com 2 e 3 camadas ocultas, utilizando o método de Powell-Beale apresentaram melhor desempenho, com coeficientes de correlação superiores a 0,999.

Para o Modelo II, o método de Powell-Beale se mostrou bastante eficiente, sendo as RNAs com 2 camadas as mais precisas. Nesse caso, os coeficientes de correlação foram iguais a 0,999.

6.4 Modelos Alternativos

Como mencionado na Seção 5.3, a obtenção dos Modelos III e IV foram propostos como um desdobramento dos resultados obtidos com a ACP. A proposta do Modelo III se deveu ao fato de que a ACP não indicou, em função do corte do coeficiente de correlação, que a variável 'trombose provocada/espontânea' fosse um fator importante. No entanto, do ponto de vista prático, esse é um fator com alta relevância, pois é considerado que pacientes com trombose provocada possuem um menor risco de TR, sendo o tratamento desses pacientes mais simples e rápido. É normal que a ACP não indique todos os fatores

relevantes do ponto de vista prático, uma vez que é uma técnica que analisa de modo puramente matemático o conjunto de dados considerado. Por fim, a proposta do Modelo IV se deve ao fato de que os exames para a determinação dos níveis de proteína C, S e antitrombina possuem alto custo e, se possível, evitados de serem realizados.

Em ambos os casos, o método de Powell-Beale foi utilizado para ajustar os parâmetros das RNAs. Os resultados para o Modelo III mostraram que uma RNA com três camadas ocultas são mais precisas, sendo os melhores resultados com coeficientes de correlação chegando a 0,999. As mesmas observações puderam ser notadas para o Modelo IV. Essa etapa do trabalho confirmou que as RNAs são uma ferramenta poderosa para a predição de doenças e suas complicações a partir de dados clínicos.

6.5 Utilização dos Modelos Obtidos

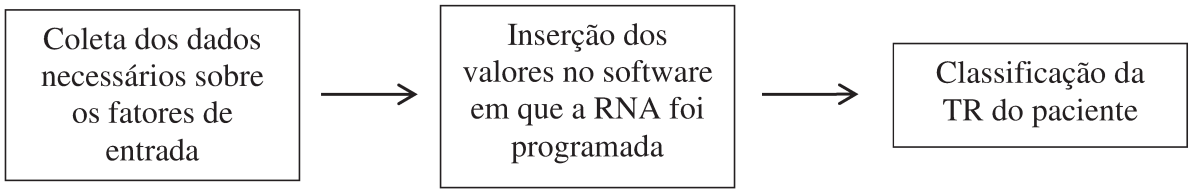
Uma RNA é uma equação matemática, conforme apresentado pelas Eqs. (3) e (4). Uma vez ajustados os pesos e os *bias* da RNA, ela pode ser facilmente programada em um software gerenciador de planilhas, utilizando uma linguagem de alto nível, ou até na forma de um aplicativo para celular. Assim, qualquer médico pode utilizar a equação facilmente na sua rotina de atendimentos.

Para se utilizar a RNA, recomenda-se a obtenção de todos os dados referentes às variáveis de entrada do modelo, necessários para o cálculo. Uma vez obtidos os valores, basta inserir no programa utilizado e efetuar o cálculo da resposta. A Figura 6.1 ilustra o procedimento de uso da RNA treinada para a predição da TR.

Vale salientar que se a RNA utilizada for a referente ao Modelo I, são necessárias as informações acerca dos 39 fatores considerados. Se a RNA for referente ao Modelo II, são necessárias as informações acerca dos 18 fatores definidos como importantes. Caso a RNA seja referente ao Modelo III, são necessárias informações sobre as variáveis do Modelo II, incluindo se a primeira trombose do paciente foi espontânea ou provocada. Por fim, caso a RNA se refira ao Modelo IV, são necessárias as informações sobre as variáveis do Modelo II, incluindo se a primeira trombose do paciente foi espontânea ou provocada, e excluindo-se os valores de proteína C, S e antitrombina. Não é necessário aplicar a ACP previamente à

utilização, visto que essa técnica foi apenas uma ferramenta de auxílio no desenvolvimento dos modelos.

Figura 6.1 – Diagrama de utilização da RNA.



7 PRIMEIRO MANUSCRITO: PRINCIPAL COMPONENT ANALYSIS ON RECURRENT VENOUS THROMBOEMBOLISM

Nesta Seção, serão apresentados os resultados referentes à determinação dos principais fatores preditivos da trombose. O texto foi escrito em inglês e é uma adaptação do manuscrito original que foi submetido ao jornal *Thrombosis Research* – já que não serão apresentados *highlights*, *abstract*. Além disso, as conclusões serão apresentadas no Capítulo 9.

7.1 Introduction

RVTE is an important subject in any medical center. Once the patient is diagnosed with a first venous thrombosis (VTE), anticoagulation therapy is the prevention method of recurrence commonly used. However, the duration of anticoagulation therapy is a difficult decision and depends on multiple factors that interact with each other and need to be evaluated individually (Hull *et al.*, 1979; Heit *et al.*, 2000; Farzamnia *et al.*, 2011; Guyatt *et al.*, 2012; Fahrni *et al.*, 2015).

The duration of anticoagulation therapy is a difficult decision. This is done by comparing the probability of RVTE against the probability of bleeding using score methods developed with statistical analysis of patient data. Dash (Tosetto *et al.*, 2012), Vienna (Eichinger *et al.*, 2010), and Men and HERDOO-2 (Rodger *et al.*, 2008) are three scores used to decide about anticoagulant prophylaxis by the risk of recurrence and each one takes different factors in consideration. Recent reviews showed that these score models have a strong limitation when predicting patients with low risk of RVTE and still lacks of a complete validation (Kyrle; Eichinger, 2012; Ensor *et al.*, 2016).

Patients with provoked VTE are also not included in prediction risk. There are some studies indicating that some patients with provoked VTE could benefit of prolonged prophylaxis (Hansson *et al.*, 2000; Investigators, 2010; Agnelli *et al.*, 2013) and also that they have the same risk of recurrence when comparing with those with a previous unprovoked event (Cosmi *et al.*, 2011). The scores also does not include patients with antiphospholipid antibody syndrome, cancer, natural anticoagulant activity deficiencies, etc.(Kyrle; Eichinger, 2012; Ensor *et al.*, 2016). It is important the development of prediction scores that include

those individuals in the calculation. So, to overcome these issues, sophisticated multivariate statistical techniques could be used to identify the main risk factors of RVTE for a wide range of patients. Finally, these main risks can be used in new prediction models.

PCA is a multivariate statistical method that identifies patterns and classifies the factors that influences a given phenomenon. It is a technique widely used to identify patterns in the medical field (Okin *et al.*, 2002; Kutcher *et al.*, 2013; Chin *et al.*, 2014; Thorpe *et al.*, 2016; Vavougios *et al.*, 2016). Mathematically, PCA is a decomposition of a set of correlated variables into a set of uncorrelated variables, named as Principal Components (PCs), which are organized by descending order of variance. PCA can also be used to reduce dimensionality, by cutting of the PCs that are less important (less variance) than the original data. The remaining PCs are useful to develop new models or the phenomenon, and their loadings can be used to calculate the contribution of all the factors in each PC.

The aim of this study was to collect clinical data of several patients diagnosed with a first thrombotic episode and apply PCA to identify the predictor factors for RVTE. To obtain a more comprehensive view of the risk factors, we included patients with provoked and unprovoked VTE, antiphospholipid antibody syndrome, cancer, and natural anticoagulant activity deficiencies.

7.2 Methods

7.2.1 Population

In this study, 307 patients with a first acquired or provoked VTE, assisted at outpatient clinic of Hemocentro Unicamp between January 2009 and August 2016 were followed. Acquired risk factors for VTE were pregnancy and postpartum, hormone therapy for contraception or hormonal replacement, surgery, trauma, air travel, cancer, hereditary thrombophilia (Protein C, protein S, antithrombin levels, FV Leiden mutation, G20210A prothrombin gene mutation), and antiphospholipid antibodies (lupus anticoagulant and anticardiolipin antibodies). Exclusion criteria were thrombosis that occurred in other sites than pulmonary, lower limbs or central nervous system, and age under 18 years.

Table 1 shows all the evaluated factors, which included clinical (age, sex, body mass index, tabagism, arterial hypertension, Diabetes mellitus), and thrombotic (thrombosis

site: pulmonary embolism, proximal or distal lower limb, central nervous system; provoked or unprovoked; existence of anticoagulation therapy; time of anticoagulation therapy) data, acquired and inherited risk factors for thrombosis, laboratorial parameters (hemogram, creatinine, total, HDL and LDL cholesterol, triglicerydes, glucose, C-reactive protein, D-dimer) and residual vein thrombosis detected by Ultrasound Doppler. All blood samples were collected between the VTE and RVTE events, in a mean time of 12 months after the first episode.

Table 1 – Factors considered in the present study.

Sex (men/women)	G20210A prothrombin gene mutation (No/Yes)	Red blood cell (RBC) count (μL^{-1})	Fasting blood glucose ($mg.dL^{-1}$)
Age (years)	Factor VIII ($IU.dL^{-1}$)	Hemoglobin (HB) ($g.dL^{-1}$)	Creatinine ($mg.dL^{-1}$)
Pulmonary embolism (No/Yes)	Protein C level ($mg.dL^{-1}$)	Hematocrit (HCT) (%)	C-reactive protein ($mg.dL^{-1}$)
Side of inferior member (left/right)	Protein S level ($mg.dL^{-1}$)	Red Blood Cell Distribution (RDW) (%)	Arterial hypertension (No/Yes)
Level of VTE in leg (distal/ proximal)	Antithrombin level ($mg.dL^{-1}$)	Platelet (PLT) (μL^{-1})	Diabetes mellitus (No/Yes)
Provoked or unprovoked	Antiphospholipid syndrome (No/Yes)	Mean Platelet Volume (MVP) (fL)	Dyslipidemia (No/Yes)
Anticoagulation time (months)	Body mass index (BMI)	Total cholesterol ($mg.dL^{-1}$)	Renal failure (No/Yes)
Anticoagulant use (No/Yes)	Tabagism (No/Yes)	High-density Lipoprotein (HDL) ($mg.dL^{-1}$)	Cancer (No/Yes)
D-dimer ($ng.mL^{-1}$)	Hormonal therapy (for women) (No/Yes)	Low-density Lipoprotein (LDL) ($mg.dL^{-1}$)	Residual venous thrombus on US Doppler (No/Yes)
FV Leiden (No/Yes)	White blood cell (WBC) count (μL^{-1})	Triglycerides ($mg.dL^{-1}$)	

The 39 factors presented in Table 1 were used to perform the PCA. Since there are some input variables that are not numerical data, they were converted to a numerical form, as shown in Table 2. All numerical variables were used as is.

Table 2 - ANN inputs for non-numerical variables.

Factor	ANN input value
Sex	Male = 1; Female = -1
Pulmonary embolism	Yes = 1; No = -1
Site of thrombosis	Right inferior member = 1 Left inferior member = -1
Proximal or distal lower limb thrombosis	Proximal = 1; Distal = -1
Provoked or unprovoked	Provoked = 1; Unprovoked = -1
Anticoagulant use	Yes = 1; No = -1
FV Leiden	Negative = 1; Heterozygous = -1
G20210A prothrombin gene mutation	Negative = 1; Heterozygous = -1
Antiphospholipid syndrome	Negative = 1; Positive = -1
Tabagism, hormonal therapy, arterial hypertension, diabetes mellitus, dyslipidemia, cancer, residual venous thrombus on US doppler	Yes = 1; No = -1

7.2.2 Principal Component Analysis

To calculate the PCs from an original data matrix **X**, it is necessary to eliminate data heterogeneity. This can be done by centering the mean in zero and the variance in one, for each variable from **X**. Eq. (20) is used for this purpose:

$$Z_{i,j} = \sigma_j^{-1/2}(X_{i,j} - \mu_j); i = 1,\ n; j = i, m \tag{20}$$

where: $X_{i,j}$ is the *ith* data of variable *j*, $Z_{i,j}$ is the stardardized variable of $X_{i,j}$, and μ_j and σ_j are the mean value and the variance of the variable *j*, respectively.

Once the matrix **Z** is obtained the variance-covariance matrix can be obtained. The PCs can be generated by calculating the eigenvector and the eigenvalues for the matrix **Z**. The importance of each PC can be determined by ordering the eigenvalues in a descending order, corresponding to the descending order of variance.

In this study, PCA was used to determine the main factors for RVTE. We considered only the PCs which eigenvalue is greater than 1. To compute the importance ofeach factor Z_i in a PC_j, the correlation, $\rho_{PC_jZ_i}$, was calculated by using Eq. (21):

$$\rho_{Z_iPC_j} = a_{ij}\sqrt{\lambda_i}; i,\ j = 1,\ 2,\ldots,n \tag{21}$$

where: a_{ij} is eigenvector value of the i^{th} stardardized variable Z with respect to the j^{th} PCs, and λ_i is the eigenvalue of the i^{th} stardardized variable Z.

Only those variables with $\left|\rho_{PC_jZ_i}\right| > 0.5$ were considered as important factor.

7.3 Results

From all 307 patients, 72 were excluded because they presented a first VTE in other sites than pulmonary, lower limbs or central nervous system, and age under 18 years. Therefore, 235 patients were included whereas 74 were men (31.4 %) and 161 were women, age >18 and <82 years.

Table 3 summarizes the principal characteristics of the thrombotic episodes. 49 patients (20.8 %) presented recurrent VTE. The first episode were more predominant in the left leg (45.7 %). 36 % of the patients presented pulmonary embolism and only 2.1 % presented thrombosis in the central nervous system. When analyzing RVTE events, the percentages were almost the same: 46.0 %, 30.0 %, and 2.0 % for patients who presented VTE in the left, right, and both legs, respectively. 22 % presented TEP as RVTE, as well as 2.0 % in the central nervous system. 39 % of all patients presented a provoked first thrombosis and 40 % of these patients had RVTE, corresponding to 16 % of all recurrences. This is an interesting result, since recurrence in patients with a provoked VTE is expected to be lower. Cancer was presented in 13 patients. 3 of them had provoked VTE and only 1 presented RVTE.

These RVTE episodes were not in the same local for all patients. Table 4 shows the relation the VTE/RVTE local for all patients who presented RVTE. The RVTE events were not in the same local for all patients. Table 4 shows the relation the VTE/RVTE local for all patients who presented RVTE. It can be seen that the thrombotic events repeat more

frequently in the same local for patients who presented VTE in the left leg: from 24 VTE patients that presented VTE in the left leg, 16 presented RVTE in the same local. For other events, this observation is not the same.

Table 3 – Main characteristics of the thrombotic events

Time between thrombotic episodes	23.75 ± 19.12 months	
Time between the end of anticoagulation therapy and thrombotic episodes	15.87 ± 14.89 months	
	VTE	RVTE
Men	74	20
Women	161	29
Local		
Left leg	107	23
Right leg	71	15
Pulmonary embolism (exclusive)	42	9
Pulmonary embolism	85	11
Central nervous system	5	1
Both legs	6	1

Table 4 – Relation between VTE and RVTE events.

VTE \ RVTE	Left leg	Right leg	Both legs	Central nervous system	TEP	Total RVTE
Left leg	16	4	0	2	1	23
Right leg	6	6	1	2	0	15
Both legs	0	0	1	0	0	1
Central nervous system	0	1	0	0	0	1
TEP	2	3	1	1	2	9
Total VTE	24	14	3	5	3	49

In this study, we used all gathered data to calculate the eigenvalues and the eigenvectors required to obtain the PCs, and then using Eq. (21), to determine the main factors for recurrent VTE. The eigenvalues obtained and the cumulated variance for each PC is shown in Figure 1. The correlations calculated by Eq. (21) between the factors and the 13

PCs considered are presented in the Supporting Information (Table S1). A factor can be considered highly correlated with a PC, therefore an important factor, when its correlation modulus is higher than 0.5. The correlations for the other PCs (14-39) were not presented since they were discarded. The main factors for RVTE found in this work are presented in Table 5 in order of importante.

Table 5 – Main factors for RVTE determined by PCA.

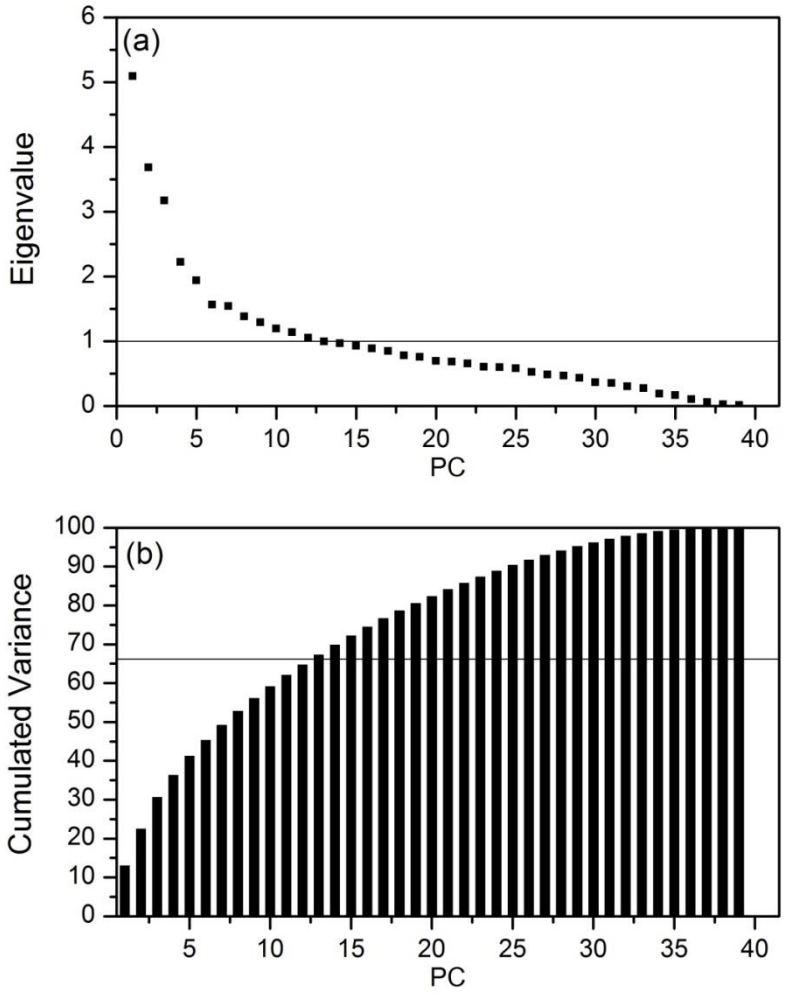
PC	Correlation
PC 1	
Total Cholesterol	0.765
LDL	0.706
HDL	0.658
Creatinine	0.564
Triglycerides	0.551
Glucose	0.520
HCT	0.520
RBC	0.508
PC 2	
AT	0.712
PS	0.688
PC	0.661
Age	-0.558
PC 3	
HB	0.781
HCT	0.770
RBC	0.749
PC 5	
WBC	0.569
RDW	0.525
PC 7	
D-dimer	0.512
PC 9	
Level of VTE in leg	0.657
PC 12	
Anticoagulation time	0.522

The results indicated that 26 of the 39 PCs correspond to 90.37 % of the overall variance. In Figure 1a, the line y = 1 correspond to the marginal value bellow which a PC

account for less variance than the original ones. So, only the PCs with eigenvalues greater than 1 were considered. 13 PCs satisfy this condition, so we believe that they are sufficient to describe RVTE, corresponding to 67.34 % of the total variance (horizontal line in Figure 1b).

The correlations calculated by Eq. (21) between the factors and the 13 PCs considered are presented in the Supporting Information (Table S1). A factor can be considered highly correlated with a PC, therefore an important factor, when its correlation modulus is higher than 0.5. The correlations for the other PCs (14-39) were not presented since they were discarded. The main factors for RVTE found in this work are presented in Table 5 in order of importance.

Figure 1 - PCA results. (a) eigenvalues (b) cumulated variance.



The first component accounted for 13 % of overall variance, including RBC, HCT, triglycerides, glucose, creatinine, cholesterol, HDL, and LDL. The former three factors presented much higher correlations with PC1, indicating that they have strong influence in this PC.

The second component accounted for 9 % of the overall variance, including age, protein C, protein S and antithrombin levels. The third component accounted for 8 % of the overall variance, including RBC, HB, and HCT.

The other highly correlated variables were: WBC and RDW (5^{th} PC – 5.7 % of overall variance), d-dimer (7^{th} PC – 4 % of overall variance), proximal or distal lower limb thrombosis (9^{th} PC – 3.5 % of overall variance), and duration of anticoagulant treatment (12^{th} PC – 2.92 % of overall variance). The other PCs presented no correlation modulus higher then 0.5, indicating that there is no relative importance between the factors in those PCs.

Based on the results, one can relate each PC with a determined cluster of variables. As we seen, the 1^{st} PC is characterized by RBC, total cholesterol, LDL and HDL cholesterol, tryglicerides, glucose level, as well creatinine level. The 2^{nd} PC relates with age and natural anticoagulants levels. The main factors in 3^{rd} PC are RBC parameters. The size distribution of RBC, and WBC count are the main factors in the 5^{th} PC. The others PCs presented only one factor as important, so, they are mainly characterized by them.

7.4 Discussion

It is well known that the factors for rethrombosis identified by PCA are related to thrombotic events (Hansson *et al.*, 2000; Brækkan *et al.*, 2010; Farzamnia *et al.*, 2011) as transient or persistent risk factors such as recent surgery, major trauma, pregnancy, puerperium, hormone replacement therapy and oral contraceptive use and cancer (Heit, 2012). However, since RVTE is a multifactorial disease, the exact mechanism, and interactions between these factors still needs a better elucidation.

7.4.1 First PC

The main factors given by the 1^{st} PC corroborate recent publications regarding the association of RBC, lipids, glucose and creatinine levels and recurrent VTE. From Table S1,

one can see that cholesterol, HDL and LDL had higher correlation modulus than the others. This result means that they highly influence PC1.

The role of the lipids in VTE is still uncertain and contradictory results were reported (Deguchi *et al.*, 2005; Everett *et al.*, 2009). It was shown that HDL influences the extrinsic coagulation pathway, the protein C cascade, the fibrinolysis and also reduces blood viscosity. Therefore it has antithrombotic properties. On the other hand, triglycerides have procoagulant effect, suppressing the tissue plasminogen activator activity and increasing PAI-1 (Belaj *et al.*, 2014). A recent prospective study evaluated the lipid levels in 2106 patients (Morelli *et al.*, 2017). The authors affirmed that there was no association between these lipids level and recurrent VTE. Eichinger *et al.* (2007) reported that lower levels of HDL was associated with recurrent VTE. In both studies, blood samples were collected after 3 months of discontinuation of anticoagulation. Our results showed that patients with recurrent VTE presented higher levels of total cholesterol (193.29 ± 44.25 mg/dL vs 184.90 ± 49.19 mg/dL), HDL (49.29 ± 14.91 mg/dL vs 46.41 ± 14.45 mg/dL), LDL (114.38 ± 35.49 mg/dL vs 105.56 ± 35.04 mg/dL) than those without recurrence. On the other hand, the level of triglycerides was lower (148.52 ± 60.00 mg/dL vs. 156.04 ± 152.66 mg/dL). It is important to mention that when looking at the median, it can be observed that the values for HDL are 45 mg/mL and 46 mg/mL for patients with and without RVTE, respectively. Moreover, for tryglicerides the median is 137 mg/dL and 122 mg/dL, for patients with and without RVTE, respectively. Both results indicate that the statistical distribution of these factors values is in agreement to those reported in the literature.

The RBC gained importance in thrombotic events only in the last decades, when its effects in blood rheology and endothelium interactions were taken into account as a venous thrombotic risk factor. (Alt *et al.*, 2002; Vayá; Suescun, 2013; Litvinov; Weisel, 2017). A recent review (Byrnes; Wolberg, 2017) showed that the RBC influence on recurrent thrombosis can be difficult to interpret, but that antithrombotic RBC targets can be developed. As demonstrated by Yu *et al.* (2011), a local increase of RBC concentration is directly related to an increase in local blood viscosity, which could promote thrombosis recurrence in humans. Marchioli *et al.* (2013) reported that high level of RBC in patients with polycythemia vera is a potential risk factor for thrombotic events. Another proposed mechanism suggest that RBC can interact with the vessel walls, enhance platelet aggregation

and activation, contribute to thrombin generation and bind with fibrinogen contributing with thrombus size (Byrnes; Wolberg, 2017). Our results showed that patients that presented recurrent VTE showed higher number of RBC ($4.89x10^6 \pm 0.62x10^6/\mu L$) than those without recurrence ($4.61x10^6 \pm 0.62x10^6/\mu L$), which is in agreement with cited studies.

Another factor of blood viscosity is the HCT. It was reported that an increase in HCT levels in the venous system, can decrease the blood flow due to the increase in its viscosity, thus favoring the clot formation (Wells; Merrill, 1962; Dintenfass, 1964). Also, high HCT level promotes platelet adhesion and accumulation at the subendothelial cells (Byrnes; Wolberg, 2017). Indeed, higher HCT levels were previously correlated to thrombotic events (Brækkan *et al.*, 2010). As far as we know, there is only one study that reported HCT level as a predictor of recurrent VTE. The study of Eischer *et al.* (2012) showed that patients with recurrence of VTE presented higher HCT levels (3 months after discontinuation of anticoagulation) than those without recurrence. However, higher levels of HCT correlated with recurrent VTE in women, and not in men. Our results showed that HCT level in patients with VTE recurrence were higher (42.16 % ± 4.39 %) than those without (40.70 % ± 4.95 %).

The role of glucose level in recurrence of VTE is still uncertain and as far as we know, there are no studies that analyzed its effects on it. High glucose levels can trigger the coagulation system in healthy men and in patients with diabetes mellitus (Grant, 2007; Kim *et al.*, 2014). For this reason, we choose to include this factor in the study. Our results demonstrated that patients with recurrence of VTE presented lower glucose levels than those without recurrent episodes (93.12 ± 21.27 mg/dL *vs.* 105.23 ± 51.17 mg/dL, respectively). Besides this is not an expected result, recurrence is multifactorial and the interaction with other factors could influence the thrombus formation.

The role of creatinine level on RVTE is still unknown. Shlipak *et al.* (2003) reported that higher creatinine levels is accompanied with the increase in factor VII, factor VIII, and D-dimer, which can lead to VTE events. Our results showed that patients with higher levels of creatinine have RVTE episodes against those with lower creatinine levels (0.86 ± 0.23 mg/dL *vs.* 0.83 ± 0.31 mg/dL).

7.4.2 Second PC

Age and coagulation proteins are the main identifiers of 2nd PC and, as shows Table S1, their importance on the PC is similar. Age is well known as an intrinsic factor for thrombosis. Several researches indicated that the thrombosis affects 0.01 % of the population before the age of 40, and 0.7 % of the population between the ages of 45-55. In addition, the morbidity is higher as age increases (Silverstein *et al.*, 1998; Tsai *et al.*, 2002; Fahrni *et al.*, 2015). Hansson *et al.* (2000) showed that age was not a significant factor. Farzamnia *et al.* (2011) did not found a direct relation between age and a higher risk of recurrence. White *et al.* (1998) reported that recurrence was more common in younger patients, while Beyth *et al.* (1995) found that this was observed in patients aged 65 and younger. On the other hand, Heit *et al.* (2000), Eichinger *et al.* (2007) and Galanaud *et al.* (2014) reported that increasing age was related to a higher risk of recurrence. Our results showed that patients with recurrent VTE presented similar average age (43.40 ± 14.82 years) against (44.05 ± 16.07 years) those without recurrence.

Deficiency of natural anticoagulants can trigger hypercoagulability and a thrombotic event (Heeb *et al.*, 1994; Pabinger; Schneider, 1996; Esmon, 2000; Rosendaal; Reitsma, 2009; Kim *et al.*, 2014). Rosendaal and Reitsmat (Rosendaal; Reitsma, 2009) reported that these deficiencies are a strong risk factor for VTE. However, the authors affirm that this state is not necessarily true for recurrence of VTE. De Stefano *et al.* (2006) indicated that patients with these deficiencies have an increased risk for recurrent VTE. Our results showed that antithrombin, protein C, and protein S activity were important factors as patients with recurrent VTE presented lower levels of these proteins when compared to those without recurrence (Protein C: 106.26 ± 28.39 $mg.dL^{-1}$; Protein S: 90.69 ± 23.83 $mg.dL^{-1}$; antithrombin: 105.69 ± 16.11 $mg.dL^{-1}$ vs. Protein C: 117.87 ± 23.60 $mg.dL^{-1}$; Protein S: 92.41 ± 23.44 $mg.dL^{-1}$; antithrombin: 106.22 ± 15.57 $mg.dL^{-1}$, respectively). These results are very interesting as even without a classical deficiency, we showed that lower levels are important for RVTE. Our findings are in agreement Xu *et al.* (2010), whom proposed a mathematical model for the coagulation cascade and showed that lower levels of Protein C is associated with thrombus formation.

7.4.3 Third PC

The 3rd PC is characterized by the RBC parameters: RBC, HB, and HCT. RBC and HCT also were the important factors in the 1st PC. However, their importance in 3rd PC is higher. As we previously discussed, RBC and HCT can be associated with VTE and recurrence. Brækkan *et al.* (2010) found a directly relation between HB at the admission and VTE: higher the HB levels, higher was the risk of VTE. Our results showed that patients with recurrence presented similar level of HB (13.99 ± 1.44 g/dL) when compared with those without (13.54 ± 1.77 g/dL). As far as we know, there is no study that evaluated the influence of HB levels in RVTE. However it is also known that HB molecules released from damaged RBCs enchance platelet activation and aggregation (Byrnes; Wolberg, 2017).

7.4.4 Fifth PC

RDW and WBC are the major descriptors of 5th PC. The leukocyte number and the red blood cells distribution width have been related with thrombotic events (Lopresti *et al.*, 2000; Gangat *et al.*, 2007; Bucciarelli *et al.*, 2015). However, their role in recurrence of VTE is still not well established. De Stefano *et al.* (2008) showed that high WBC count at the time of the first thrombosis are correlated with recurrence of VTE in patients < 60 years. Saraiva *et al.* (2015) concluded that an increased monocyte count are an important predictor for RVTE. Carobbio *et al.* (2007) reported that, in patients with essential thrombocythemia, high WBC count prior to the thrombotic event is related to VTE recurrence. On the other hand, in the study of Rezende *et al.* (2013) reported that those with higher risk of VTE presented lower levels of WBC and higher levels of RDW after the VTE. The authors also suggested that other studies should be performed to investigate the relationship between these variables with VTE recurrence. Our results showed that patients with VTE recurrence presented lower levels of WBC ($6.89.10^3 \pm 2.69.10^3$/μL) and higher levels of RDW (14.17 ± 1.96 %) than those without recurrence (WBC: $7.94.10^3 \pm 3.86.10^3$/μL – RDW: 13.59 ± 1.49 %).

7.4.5 Remaining PCs

D-dimer, proximal or distal lower limb thrombosis, and duration of anticoagulant treatment are the descriptors of the 7th, 9th and 12th PCs, respectively. D-dimer is the most

important predictor factor for VTE and its recurrence and several studies about its influence can be found on the literature (Palareti *et al.*, 2002; Palareti *et al.*, 2003; Verhovsek *et al.*, 2008; Lippi *et al.*, 2014). Eichinger *et al.* (2003) found that 13 % of the total patients presented thrombosis recurrence and that these patients presented higher d-dimer levels 3 weeks after discontinuation of anticoagulation than those without recurrence. In a recent study, Bjøri *et al.* (2017) reported that those with d-dimer level less than 1500 $ng.mL^{-1}$ at diagnosis were associated with a low recurrence risk. Palareti *et al.* (2014) also reported that a higher level of d-dimer after 3 months of anticoagulation therapy increases the probability of recurrence. Our results, showed that d-dimer level in patients with recurrence are in agreement with all the studies cited (with RVTE: 839.13 ± 1101.97 $ng.mL^{-1}$; without RVTE: 672.68 ± 981.58 $ng.mL^{-1}$).

According to Galanaud *et al.* (2014), the influence of the level of VTE on the leg for RVTE is important. The authors concluded that those with proximal VTE presented more recurrent thrombotic events than those with distal VTE. Boutitie *et al.* (2011) reported that recurrence was higher in those with proximal VTE. Hansson *et al.* (2000) reported similar results. Our results showed that 83 % of patients with proximal VTE presented recurrence against 16 % of those with distal VTE, in agreement with the literature.

Time of anticoagulation therapy is one of the major factors that can lead or prevent the recurrence of VTE (Franco *et al.*, 2017). This is a subject of debate and its recommendations can be found in the American College of Chest Physicians (ACCP) Guidelines (Kearon *et al.*, 2016). Boutitie *et al.* (2011) reported that patients treated for 6 months or more presented lower RVTE events than those treated during 3 months. Also, that patients treated for 1 or 1,5 months presented higher risk of RVTE than those treated for 3 months. In a recent study, Agnelli *et al.* (2013) evaluated the extended anticoagulation therapy with apixaban and reported that the extended therapy reduced the risk of recurrence without increasing the risk of bleeding. Similar results are reported in (Investigators, 2010). Studies that analyze the time of anticoagulant after a first recurrence can also be found in literature (Schulman *et al.*, 1997; Van Der Hulle *et al.*, 2015). Our results showed that patients with RVTE were treated during a longer time than those without recurrence (12.72 ± 18.51 months vs 9.88 ± 11.50 months). This result was somewhat unexpected, since patients with less time of anticoagulation therapy is supposed to have higher probability of RVTE.

However, when comparing only patients with RVTE (data not shown), it could be observed that RVTE occurred mainly in patients treated during shorter periods.

7.4.6 Final Considerations

The results obtained in this work showed that RVTE, a priori, can be predicted using clinical parameters and data from the previous thrombosis. It is important to mention that some variables were expected to be within the main factors, but they were not indicated by PCA, since only variables with correlation above 0.500 usually are considered as important.

The first one is the provoked/unprovoked VTE factor. The results of this work showed that 16 % of all patients with RVTE had a first provoked VTE. In daily basis, patients with provoked VTE usually are treated during 3-6 months and then the anticoagulation therapy is ceased. However, a different decision could avoid such a high recurrence rate. This factor is reported in Table S1 with a correlation of 0.493 in PC 4, which is very close to 0.500. Due its importance to RVTE, this factor could be included as input in a prediction model. Cancer is another variable expected to be in the main factors list. However, only 13 patients of 235 presented this desease. Therefore, its influence is too small when comparing to the other 38 variables. This can be confirmed by its correlation coefficients in Table S1, which are smaller than 0.360. A future study should address only these patients.

The limitations for this study are mainly due the sample size, which is small considering the eligible population. However, for determination of the main factors using PCA, it was considered to be sufficient. Another factor that was not directly considered here is the time when the blood samples were collected. According to Rezende *et al.* (2013), the values of the factors included in the proposed models do not fluctuate significantly with time. Thus, its influence in the dataset could be despised. Finally, to perform a study with another population is important for external validation.

In summary, PCA showed that the important predictors of recurrence are in agreement with previous studies and are simple to obtain in clinical routine. These findings suggest that PCA is an important tool in clinical practice to improve the stratification of patients after VTE. Future work should consider another population for validation propurses.

7.5 Supporting Information

Table S1 – Correlation between PCs and original variables.

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13
Sex	0,133	-0,216	0,346	-0,450	-0,211	-0,251	0,145	-0,057	0,039	-0,076	0,072	0,172	-0,050
Age	0,151	**-0,558**	0,057	-0,444	0,009	0,240	0,020	0,082	0,007	0,219	-0,009	-0,112	0,189
PE	-0,030	-0,177	0,000	-0,053	0,281	0,183	0,470	-0,026	-0,313	-0,229	0,330	-0,128	0,005
Side of thrombosis	-0,062	-0,161	-0,008	-0,063	-0,012	0,350	0,192	-0,264	0,175	-0,176	0,408	-0,017	-0,338
Distal/Proximal	0,067	0,118	0,104	-0,044	-0,036	0,068	-0,125	-0,320	**0,657**	0,242	-0,130	-0,040	-0,011
Provoked/unprovoked	-0,283	0,156	-0,234	**0,493**	0,100	0,057	-0,195	-0,080	0,078	-0,015	0,314	0,257	0,036
Treatment time	0,007	-0,135	0,139	-0,003	-0,342	0,255	-0,184	-0,212	-0,107	0,083	-0,097	**0,522**	-0,309
Anticoagulant use	0,107	0,030	-0,083	0,271	0,104	0,138	0,205	0,155	-0,367	0,497	-0,164	0,300	0,005
D-dimer	0,133	-0,187	0,012	0,067	0,237	0,062	**0,512**	-0,239	0,064	-0,196	-0,261	0,086	-0,001
FV Leiden	0,122	0,498	-0,168	-0,140	0,118	-0,222	-0,045	0,380	-0,140	0,276	0,098	-0,082	-0,122
G20210A mutation	0,281	0,434	-0,143	-0,123	0,139	-0,228	0,169	0,321	0,158	0,167	0,191	-0,060	-0,111
FVIII	0,413	0,257	-0,208	-0,174	-0,049	0,037	-0,151	-0,216	-0,237	0,006	-0,064	-0,147	-0,210
PC	0,433	**0,688**	-0,169	-0,341	0,067	0,025	0,124	-0,173	-0,032	-0,081	-0,029	0,043	0,050
PS	0,394	**0,661**	-0,190	-0,310	0,105	-0,016	0,101	-0,167	-0,009	-0,069	-0,030	0,165	0,055
AT	0,375	**0,712**	-0,230	-0,288	0,043	0,018	0,117	-0,184	-0,042	-0,099	-0,105	0,129	0,041
SAF	-0,027	-0,278	0,107	0,190	-0,066	-0,353	0,153	-0,030	-0,288	-0,125	-0,195	0,099	0,174
BMI	0,248	0,106	-0,163	-0,008	-0,184	0,425	-0,072	0,001	0,096	0,040	0,256	0,007	0,463
Tabagism	-0,048	-0,244	0,204	-0,352	0,169	-0,401	-0,241	-0,088	-0,056	0,049	0,066	-0,081	-0,113
Hormonal therapy	-0,271	0,394	-0,211	0,424	0,025	0,065	-0,095	0,126	0,009	-0,073	0,203	0,139	-0,142
WBC	0,400	-0,135	0,164	0,230	**0,569**	-0,005	0,069	-0,094	0,094	-0,176	-0,154	0,071	0,030
RBC	**0,508**	0,205	**0,749**	0,136	-0,124	0,083	-0,061	0,018	-0,064	-0,005	0,119	-0,065	-0,081
HB	0,487	0,205	**0,781**	0,080	-0,133	0,046	-0,053	0,028	-0,076	0,002	0,098	-0,080	-0,040
HCT	**0,520**	0,186	**0,770**	0,093	-0,119	0,065	-0,072	0,036	-0,070	0,023	0,100	-0,076	-0,045
RDW	0,330	-0,225	0,206	0,174	**0,525**	0,070	0,000	0,116	0,048	-0,011	-0,099	-0,037	-0,279
PLT	0,384	0,147	0,087	0,276	0,270	0,073	-0,359	-0,093	-0,001	-0,150	-0,319	-0,098	0,184
MVP	0,421	0,095	0,300	0,218	0,376	0,018	0,051	0,163	0,152	0,070	0,181	0,266	0,219
Total cholesterol	**0,765**	-0,278	-0,333	0,233	-0,305	-0,113	0,076	-0,021	-0,003	-0,024	0,027	-0,079	-0,029
HDL	**0,658**	-0,186	-0,339	0,274	-0,236	0,017	-0,049	-0,015	-0,055	0,050	-0,021	-0,161	-0,025
LDL	**0,706**	-0,244	-0,321	0,196	-0,337	-0,157	0,066	-0,005	0,025	-0,075	0,003	-0,010	-0,042
Triglycerides	**0,551**	-0,286	-0,232	0,078	-0,072	-0,064	0,165	-0,046	0,066	0,107	0,143	-0,088	-0,011
Glucose	**0,520**	-0,192	-0,430	0,113	0,150	0,069	-0,213	-0,013	0,086	0,034	0,005	-0,150	-0,042
Creatinine	**0,564**	-0,283	0,023	-0,090	-0,048	-0,097	0,028	-0,139	-0,093	0,193	0,123	0,263	0,162
C-reactive protein	0,118	-0,177	-0,233	0,032	0,437	-0,134	-0,167	-0,356	-0,185	0,221	0,210	-0,062	-0,227
HAS	0,264	-0,386	-0,139	-0,382	0,117	0,286	-0,048	0,171	-0,053	0,096	-0,042	0,262	-0,095
Diabetes	0,275	-0,197	-0,138	-0,310	0,154	0,240	-0,281	0,376	0,153	-0,135	-0,192	0,097	-0,185
Dyslipidemia	0,256	-0,278	-0,085	-0,227	0,107	-0,028	-0,003	0,391	0,168	-0,256	0,163	0,120	0,105
Renal failure	0,228	-0,010	-0,122	0,069	-0,161	-0,401	-0,198	0,141	0,060	-0,442	0,074	0,291	-0,087
Cancer	-0,036	-0,256	0,043	-0,193	0,283	-0,323	-0,356	-0,325	-0,046	0,088	0,268	0,125	0,253
Residual thombus	-0,016	-0,034	0,090	0,154	-0,034	-0,335	0,419	-0,003	0,455	0,302	-0,042	0,100	-0,098

8 SEGUNDO MANUSCRITO: ARTIFICIAL NEURAL NETWORKS FOR PREDICTION OF RECURRENT VENOUS THROMBOEMBOLISM

Nesta Seção, serão apresentados os resultados referentes à obtenção dos modelos neurais para predição da TR. O texto foi escrito em inglês e é uma adaptação do manuscrito original que foi submetido ao jornal *Artificial Intelligence in Medicine* – já que não serão apresentados *highlights*, *abstract*. Além disso, as conclusões serão apresentadas no Capítulo 9.

8.1 Introduction

Venous thromboembolism (VTE) is a multiple factor disease and anticoagulation therapy is its default treatment (Kearon *et al.*, 2008). However, once it starts, the question that has to be answered is for how long patients should be treated. General guidelines usually recommend at least three months of treatment or indefinite anticoagulation therapy for patients presenting low-risk of bleeding and moderate to high risk of recurrence (Kearon *et al.*, 2008). However, this subject still is controversial.

To overcome this problem, statistical models were proposed to calculate the risk of a patient develop recurrent venous thromboembolism (RVTE). As far as is known from published literature, there are available three different methods: Dash (Tosetto *et al.*, 2012), Vienna (Eichinger *et al.*, 2010; Eichinger *et al.*, 2014), and HERDOO2 (Rodger *et al.*, 2008). Each one takes different prediction factors in consideration and can be used only with patients that had a previous unprovoked VTE.

Recent reviews showed that these score models have a significant limitation when predicting patients with low risk of RVTE and still lacks of a complete validation (Kyrle; Eichinger, 2012; Ensor *et al.*, 2016). To overcome these issues, one suggests that more factors should be included in the modeling equation, as well as the validation in different populations. However, this could be a hard task, since the relation among clinical factors is difficult to model.

As far as mathematical tools are concerned, Artificial Neural Networks (ANNs) are a powerful tool that can relate dependent to independent variables without a prior

knowledge about the relation among them and able to deal with noise data. This can be done knowing which independent factors are important for the phenomenon and presenting the information to the ANN, so it can learn about it.

ANNs are known as universal function approximators (Hornik *et al.*, 1989; Haykin, 2005). They consist in a set of computational units (or artificial neurons) that are interconnected and organized into layers. An ANN consists of input layer, hidden layers (one or more depending upon the requirements) and the output layer. The number of neurons in the input and output layer is fixed and equal to the number of inputs and outputs, respectively. Figure 2 shows the general architecture of a feedforward ANN with one hidden layer.

The information process in an ANN begins relating the input (x_i) of each neuron to a synaptic weight (w_{ij}) that assesses this entry influence on the output of this neuron. Also, a bias parameter (b_j) is assigned to each ANN neuron and its purpose is to obtain a more versatile model. The sum of all these parameters makes the neuron activation (Eq. (22)) and the answer of the input stimulus is obtained applying an activation function to the neuron activation, generally the sigmoid or the hyperbolic tangent function. This procedure extends to the exiting layer of the network, where the neuron answer is the dependent variable of the problem (y_i) (Haykin, 2005).

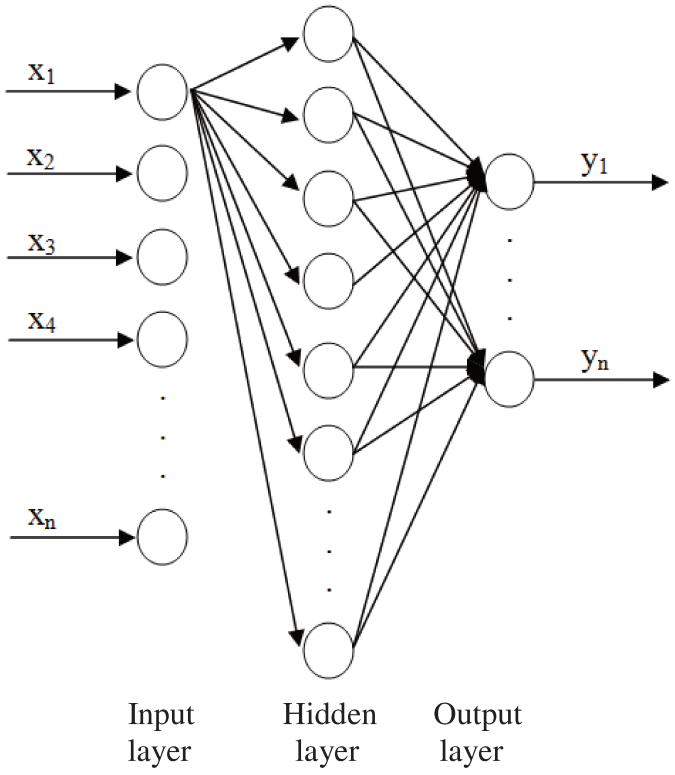
$$\alpha_j = \sum_{i=1}^{Z} w_{ij} x_i + b_j \tag{22}$$

where α_j is the neuron activation and Z is the total number of inputs for the neuron j.

Recently, ANNs have been used with success as clinical decision support systems to predict several diseases, including: dengue, chronic myeloid leukemia, acute cerebral ischemia, portosplenomesenteric venous thrombosis, portal vein thrombosis, heart diseases, asthma, etc (Badnjević *et al.*, 2016; Sasaki *et al.*, 2016; Abedi *et al.*, 2017; Babu *et al.*, 2017; Fei *et al.*, 2017a; Fei *et al.*, 2017b; Saha; Mandal, 2017). A recent review from Parveen *et al.* (2016) shows several applications and advances concerning the application of ANNs in medical field.

When working with ANNs, data reduction is a common practice, especially when a complex phenomenon with several variables is the problem to be treated. Principal Component Analysis (PCA) is a multivariate statistical method that identifies patterns and classifies the factors that influences a given phenomenon. PCA can also be used to reduce dimensionality, by cutting of the PCs that are less important (less variance) than the original data. The technique is widely used to identify patterns in the medical field (Okin *et al.*, 2002; Kutcher *et al.*, 2013; Chin *et al.*, 2014; Thorpe *et al.*, 2016; Vavougios *et al.*, 2016), and was also used to identify the main predictors of RVTE (Martins *et al.*, 2017).

Figure 2 - General one hidden layer ANN structure.



Bearing all this in mind, the aim of this work is to obtain an ANN to predict RVTE in patients with previous provoked and unprovoked VTE using clinical factors as inputs. Also, to associate PCA with ANN to obtain a model with less inputs and similar accuracy. The main contribution of this work is to show that the association of PCA with

artificial intelligence is a powerful strategy to treat complex phenomenon such as RTVE. Also, that the models obtained are suitable to predict RVTE with high accuracy.

8.2 Methods

When working with neural networks, there are three main challenges: i) to find the optimal number of hidden layers, ii) to find the optimal number of neurons in each hidden layer, and iii) to find suitable weights for each ANN neuron to the phenomenon on issue. This process is made on a step called network training and it is on this step that the network learning is evaluated. The training step is an optimization problem that depends on the initial guess and on the minimization algorithm chosen. Once the ANN is trained, the next step consists in testing the model using new data. This step is called validation. If the errors in validation step are low enough, the ANN is ready to use.

The ANN model consists in a complex equation that can be implemented in any software or program language. For example, an ANN that contains one hidden layer and W neurons in this hidden layer, based on Eq. (22), one can write the equation for the output y_n as:

$$y_n = f\left(\sum_{j=1}^{W} w_{jk} f\left(\sum_{i=1}^{Z} w_{ij} x_i + b_j\right)_j + b_k\right)_n \tag{23}$$

where: b_j e b_k are the bias, and w_{ij} e w_{jk}, are the hidden and output layer synaptic weights, respectively. So, one can write Eq. (23) in any programming software, or spreadsheet and use it in daily basis.

As part of the new procedure two ANN were proposed in this work. The first one is a complete model that includes all factors presented in Table 6 as inputs. The second ANN is a simplified model that uses as inputs some important factors determined by Principal Component Analysis in a previous work (Martins *et al.*, 2017). To develop these models, several ANN structures were trained and to confirm that it is valid we perform two validation steps (validation and test).

8.2.1 Population

All data used in this study is reported elsewhere (Martins *et al.*, 2017). 235 patients with a first VTE seen between January 2009 and August 2016 were included. Patients with provoked VTE included pregnancy, or female hormone intake, surgery, trauma, with a natural inhibitor deficiency, or lupus anticoagulant. The 235-patient study included men (75 patients) and women (161 patients). Exclusion criteria were thrombosis that occurred in other sites than pulmonary, lower limbs or central nervous system, and age under 18 years. From all patients, 49 of the 235 patients presented RVTE. In Table 6 are shown all the factors used in this work, that includes the clinical, inherited and acquired molecular ones, as well as molecular risk factors.

Table 6 – RVTE factors considered in this work.

Sex (men/women)	G20210A prothrombin gene mutation (No/Yes)	Red blood cell (RBC) count (μL^{-1})	Fasting blood glucose ($mg.dL^{-1}$)
Age (years)	Factor VIII ($IU.dL^{-1}$)	Hemoglobin (HB) ($g.dL^{-1}$)	Creatinine ($mg.dL^{-1}$)
Pulmonary embolism (No/Yes)	Protein C level ($mg.dL^{-1}$)	Hematocrit (HCT) (%)	C-reactive protein ($mg.dL^{-1}$)
Side of inferior member (left/right)	Protein S level ($mg.dL^{-1}$)	Red Blood Cell Distribution (RDW) (%)	Arterial hypertension (No/Yes)
Level of VTE in leg (distal/ proximal)	Antithrombin level ($mg.dL^{-1}$)	Platelet (PLT) (μL^{-1})	Diabetes mellitus (No/Yes)
Provoked or unprovoked	Antiphospholipid syndrome (No/Yes)	Mean Platelet Volume (MVP) (fL)	Dyslipidemia (No/Yes)
Anticoagulation time (months)	Body mass index (BMI)	Total cholesterol ($mg.dL^{-1}$)	Renal failure (No/Yes)
Anticoagulant use (No/Yes)	Tabagism (No/Yes)	High-density Lipoprotein (HDL) ($mg.dL^{-1}$)	Cancer (No/Yes)
D-dimer ($ng.mL^{-1}$)	Hormonal therapy (for women) (No/Yes)	Low-density Lipoprotein (LDL) ($mg.dL^{-1}$)	Residual venous thrombus on US Doppler (No/Yes)
FV Leiden (No/Yes)	White blood cell (WBC) count (μL^{-1})	Triglycerides ($mg.dL^{-1}$)	

8.2.2 ANN 1 (Modelo I)

For ANN 1, all gathered data was used for training, validation and test. The 39 factors presented in Table 6 were used as input variables for the model. Since there are some input variables that are not numerical data, they were converted to a numerical form, as shown in Table 7. All numerical variables were used without any conversions. The desired output was the occurrence of RVTE, which is also a non-numerical variable. It was valued as 1 if positive, and -1 if negative.

Table 7 - ANN inputs for non-numerical variables.

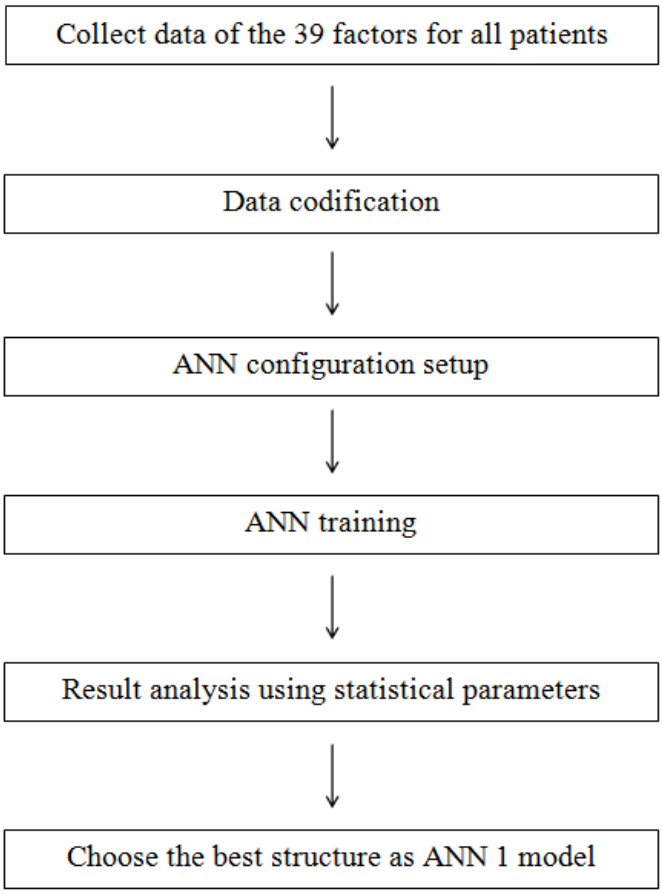
Factor	**ANN input value**
Sex	Male = 1; Female = -1
Pulmonary embolism	Yes = 1; No = -1
Site of thrombosis	Right inferior member = 1 Left inferior member = -1
Proximal or distal lower limb thrombosis	Proximal = 1; Distal = -1
Provoked or unprovoked	Provoked = 1; Unprovoked = -1
Anticoagulant use	Yes = 1; No = -1
FV Leiden	Negative = 1; Heterozygous = -1
G20210A prothrombin gene mutation	Negative = 1; Heterozygous = -1
Antiphospholipid syndrome	Negative = 1; Positive = -1
Tabagism, hormonal therapy, arterial hypertension, diabetes mellitus, dyslipidemia, cancer, residual venous thrombus on US doppler	Yes = 1; No = -1

In this work, three different optimization methods were used to train the ANNs, to know, Resilient Backpropagation (Riedmiller; Braun, 1992a), Levemberg-Marquardt (Marquardt, 1963), and Powell-Beale (Beale, 1972; Powell, 1977). All simulations were performed using the Neural Network Toolbox, from Mathworks MatLab®. The activation function in all layers was hyperbolic tangent. ANNs with one, two and three hidden layer was tested and the number of neurons varied between 10 and 40 in all layers. The objective function minimized by the methods was the mean square error given by the Eq. (24):

$$F_{OBJ} = \sum_{i=1}^{m} \frac{\left(y - y'\right)^2}{m} \tag{24}$$

which *y* is the real data value, *y'* is the ANN predicted data, and *m* is the total number of real data provided to the ANN. From all available data, 70 %, 15 %, 15 %, were used for training, validation and test steps. Data division was aleatory and performed automatic by the software. After the simulations were performed, all calculated data was compared with real ones to obtain the correlation coefficient and to evaluate the accuracy. Figure 3 shows a flow diagram of the ANN 1 training.

Figure 3 – Diagram for ANN 1 model definition



8.2.3 ANN 2 (Modelo II)

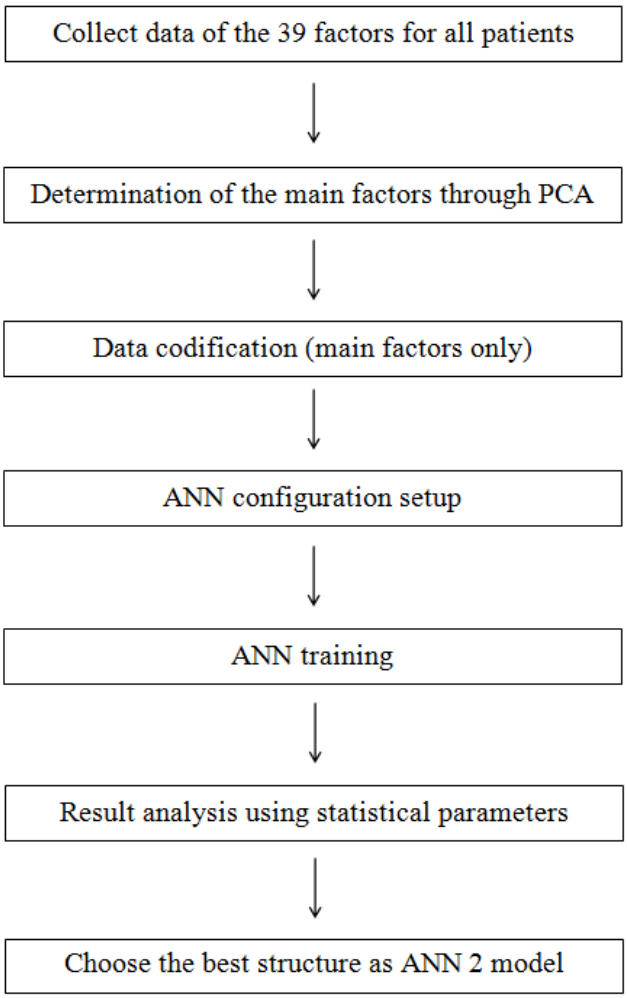
In a recent study (Martins *et al.*, 2017), it was applied Principal Component Analysis (PCA) (Hotelling, 1933b, a) in the collected data and the results indicated that for our population 18 of all 39 factors are important for RVTE prediction, as depicted in Table 8. All variables are numerical, except factor 2, which can be converted according Table 7. Since these important factors are simple to obtain, it was proposed in this work an association between the PCA and the ANN, to obtain a model with 18 important factors as inputs.

Table 8 – Input variables for ANN 2.

Age	HB
Proximal or distal lower limb thrombosis	HCT
Anticoagulant treatment duration	RDW
D-dimer	Cholesterol
Protein C	HDL
Protein S	LDL
Antithrombin	Triglycerides
WBC	Glycemia
RBC	Creatinine

The Powell-Beale method was used to train the ANNs (as shown in Results Section this methods presented better results for this application). All other setup data was the same as for ANN 1 (activation functions, F_{OBJ}, number of hidden layers and neurons, number of patients, data division) and are given in Population Section. Figure 4 shows a flow diagram of the ANN 2 training procedure.

Figure 4 – Diagram for ANN 2 model definition



8.3 Results

In this work, two models for RVTE prediction were proposed based on ANN modeling. To perform this task, clinical data of 235 patients were collected and used to train, validate and test the ANNs. In this section the results for ANN 1 and ANN 2 are presented.

8.3.1 ANN 1 (Modelo I)

The ANN 1 model consisted in networks with 39 input variables and one output, the positive/negative response for RVTE. Three optimization algorithms were used to train several structures with one, two and three hidden layers. In Tables 9, 10 and 11 are presented

the F_{OBJ} for training and validation steps, correlation coefficients for training, validation and test, as well as the number of parameters for each tested structure.

The results show that the F_{OBJ} values for training step does not follow any general tendency as the number of parameters increases. Besides, the performance of each optimization algorithm are different, as well as the accuracy of each ANN structure. In general, when comparing training F_{OBJ} value, the ANNs trained with Levemberg-Marquardt algorithm presented the worse performance among all. The best F_{OBJ} values were obtained when using Powell-Beale algorithm.

Table 9 - Results for ANNs trained with Resilient Backpropagation algorithm

Structure	F_{OBJ} Training	F_{OBJ} Validation	r^2 Training	r^2 Validation	r^2 Test	# of Parameters
39-10-1	**0,026**	**0,121**	**0,973**	**0,896**	**0,999**	**411**
39-30-1	0,222	0,289	0,779	0,764	0,779	1231
39-35-1	**0,030**	**0,050**	**0,974**	**0,944**	**0,818**	**1436**
39-5-5-1	0,159	0,174	0,836	0,824	0,696	236
39-10-5-1	0,297	0,437	0,699	0,666	0,616	461
39-15-5-1	0,359	0,464	0,684	0,661	0,382	686
39-20-5-1	0,263	0,152	0,795	0,803	0,533	911
39-30-5-1	0,185	0,152	0,829	0,866	0,817	1361
39-5-10-1	0,371	0,309	0,651	0,676	0,745	271
39-10-10-1	0,251	0,446	0,759	0,504	0,518	521
39-15-10-1	0,126	0,011	0,875	0,995	0,902	771
39-20-10-1	0,185	0,023	0,817	0,983	0,660	1021
39-30-10-1	0,196	0,233	0,899	0,789	0,803	1521

Also, the best results for each algorithm were obtained by different ANN architectures. When using Resilient Backpropagation and Levemberg-Marquardt algorithms to train the ANNs, the best results were obtained for simpler structures when comparing with the ANNs trained using Powell-Beale algorithm. However, their predictions were poor and the r^2 was very low. This behavior may explained by the differences in the weight/bias actualization performed by each algorithm, which has a direct impact in the search for the function global minimum.

Table 10 - Results for ANNs trained with Levemberg-Marquardt algorithm

Structure	F_{OBJ} Training	F_{OBJ} Validation	r^2 Training	r^2 Validation	r^2 Test	# of Parameters
39-10-1	0,057	0,003	0,950	0,994	0,999	411
39-30-1	**0,008**	**$3{,}33.10^{-5}$**	**0,999**	**0,999**	**0,892**	**1231**
39-35-1	0,157	0,178	0,869	0,811	0,702	1436
39-5-5-1	0,355	0,320	0,642	0,651	0,706	236
39-10-5-1	0,120	0,324	0,801	0,601	0,458	461
39-15-5-1	0,210	0,198	0,701	0,713	0,529	686
39-20-5-1	0,185	0,301	0,827	0,823	0,819	911
39-30-5-1	0,381	0,489	0,716	0,672	0,736	1361
39-5-10-1	0,318	0,416	0,658	0,554	0,533	271
39-10-10-1	0,668	0,231	0,808	0,650	0,131	521
39-15-10-1	0,375	0,283	0,570	0,768	0,740	771
39-20-10-1	0,341	0,411	0,691	0,565	0,643	1021
39-30-10-1	0,398	0,618	0,504	0,572	0,759	1521

Table 11 - Results for ANNs trained with Powell-Beale algorithm.

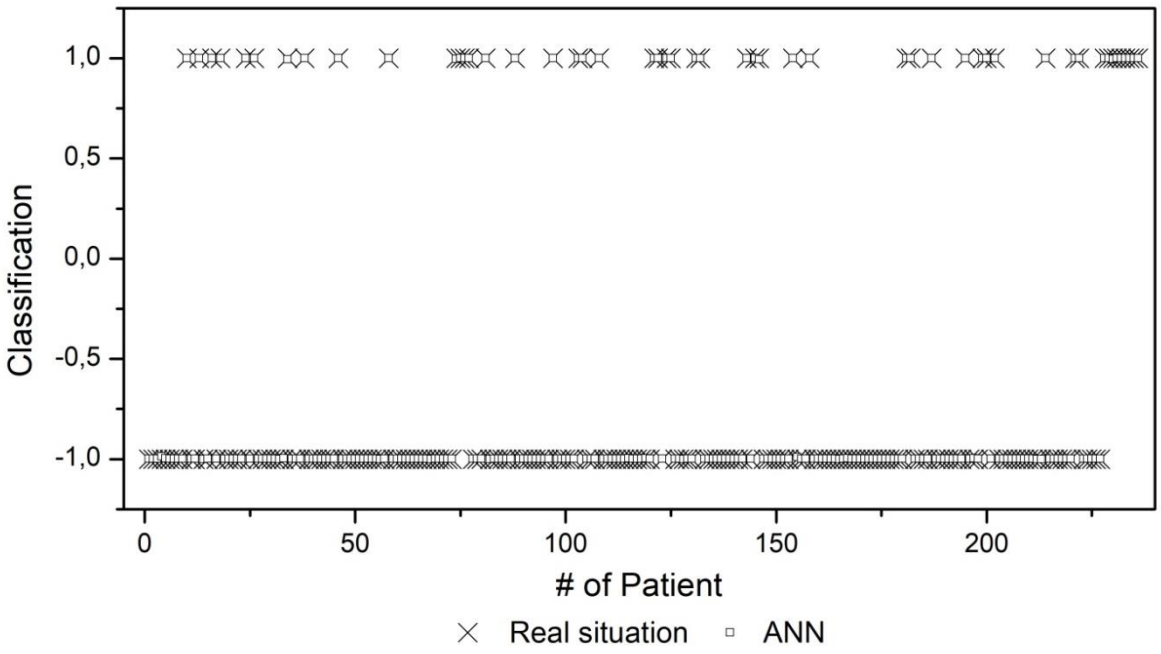
Structure	F_{OBJ} Training	F_{OBJ} Validation	r^2 Training	r^2 Validation	r^2 Test	# of Parameters
39-10-1	0,465	0,362	0,844	0,736	0,596	411
39-30-1	0,379	0,314	0,652	0,835	0,878	1231
39-35-1	0,185	0,301	0,841	0,778	0,742	1436
39-5-5-1	0,217	0,274	0,987	0,535	0,579	236
39-10-5-1	0,177	0,303	0,831	0,747	0,779	461
39-15-5-1	0,222	0,209	0,774	0,845	0,313	686
39-20-5-1	0,090	$1{,}46.10^{-5}$	0,978	0,999	0,919	911
39-30-5-1	0,154	0,332	0,851	0,845	0,89	1361
39-5-10-1	0,182	0,456	0,844	0,77	0,801	271
39-10-10-1	0,213	0,239	0,831	0,827	0,743	521
39-15-10-1	0,025	0,121	0,976	0,908	0,923	771
39-20-10-1	2.10^{-5}	$7{,}84.10^{-6}$	0,999	0,999	1	1021
39-30-10-1	0,005	0,001	0,999	0,999	0,896	1521
39-5-5-5-1	**$4{,}82.10^{-6}$**	**$1{,}46.10^{-11}$**	**0,99999**	**0,99999**	**0,99999**	**266**

The best result was obtained by the ANN 39-5-5-5-1. The F_{OBJ} values for training and validation step was the lowest obtained: in order of 10^{-6} and 10^{-11}, respectively. The r^2 obtained was very close to 1 for all stages. This structure has three hidden layers and present 265 adjustable parameters. Besides its complexity, this structure is one that presents the lower

number of parameters when comparing with others. This result appears to be due the flexibility properties that a three hidden layer ANN presents. To visualize these results, a plot of classification vs number of patient was generated and is shown in Figure 5.

In Figure 5, the points in classification 1 represent patients with RVTE and the points in classification -1 represent patients that did not presented RVTE. It is possible to see that the prediction ANN 39-5-5-5-1 is very accurate, since all the points predicted by the ANN overlapped real situation points.

Figure 5 - Comparison between ANN 1 (ANN 39-5-5-5-1) prediction and patient classification. Points for training, validation and test steps were included.



8.3.2 ANN 2 (Modelo II)

The ANN 2 model consisted in networks with 18 input variables and one output, the positive/negative response for RVTE. Due its performance, only Powell-Beale algorithm was used to train the ANN structures. Networks with one, two and three hidden layers were tested. In Table 12 are presented the F_{OBJ} and correlation coefficients for training, validation and test steps, as well as the number of parameters for each tested structure.

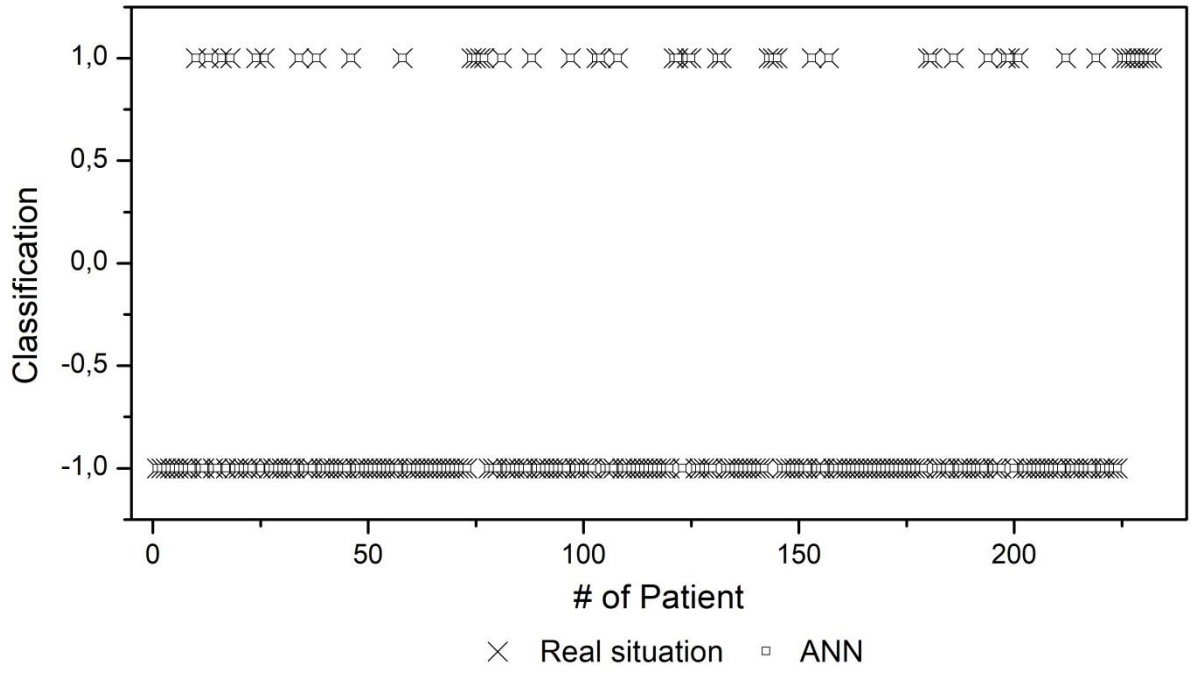
Table 12 - Results for ANNs trained with Powell-Beale algorithm.

Structure	F_{OBJ} Training	F_{OBJ} Validation	r^2 Training	r^2 Validation	r^2 Test	# of Parameters
18-10-1	0,419	0,406	0,614	0,68	0,587	201
18-30-1	0,512	0,510	0,366	0,397	0,412	601
18-35-1	0,564	0,311	0,388	0,569	0,561	701
18-5-5-1	0,367	0,338	0,617	0,593	0,351	131
18-10-5-1	0,449	0,249	0,529	0,525	0,533	251
18-15-5-1	0,564	0,702	0,289	0,227	0,115	371
18-20-5-1	0,532	0,700	0,348	0,243	0,404	491
18-30-5-1	0,541	0,287	0,372	0,415	0,476	731
18-5-10-1	0,367	0,338	0,617	0,593	0,351	166
18-10-10-1	0,467	0,922	0,403	0,389	0,429	311
18-15-10-1	0,447	0,601	0,503	0,126	0,617	456
18-20-10-1	0,398	0,788	0,514	0,284	0,422	601
18-30-10-1	**$8{,}80.10^{-4}$**	**$5{,}67.10^{-6}$**	**0,999**	**0,999**	**0,999**	**891**
18-5-5-5-1	0,059	0,12	0,957	0,912	0,861	161
18-10-5-5-1	0,035	0,121	0,97	0,908	0,999	281

The results presented in Table 12 show that the training F_{OBJ} values also did not presented any general tendency as the number of parameters increases. Besides, these values were higher in comparison with the ANN 1 models trained with the same algorithm. The best model was the structure 18-30-10-1, which has 890 adjustable parameters. The F_{OBJ} values for training and validation step in order of 10^{-4} and 10^{-6}, respectively. The r^2 obtained was very close to 1 for all steps.

It is understood that this behavior is due the reduction of variables accomplished using PCA. To model such complex phenomenon a reduced number of variables can be more difficult since the reduction of the number of parameters for the same number of neurons in hidden layers reduces the convergence space. Thus, since the number of inputs decreases, in general more neurons are needed to extract the relations for the same phenomenon. In Figure 6, which shows the plot of classification vs number of patient, it is possible to observe that the points predicted by the ANN 18-30-10-1 also overlapped all real situation points.

Figure 6 - Comparison between ANN 2 (ANN 18-30-10-1) prediction and patient classification. Points for training, validation and test steps were included.



8.4 Discussion

In this work, several ANN structures were trained using different optimization algorithms, in an attempt to obtain a new model capable to predict RVTE. Two models were proposed differing only in the number of inputs: 39 and 18, which was identified by PCA . The results showed that both ANN are capable to perform this task with high accuracy.

Several studies that evaluate RVTE predictors can be found in the literature. However, developing clinical decision rules is a difficult task because usually one analyzes one factor at a time and the number of patients is limited. To overcome these drawbacks, some methods to calculate the risk of RVTE were proposed. Dash (Tosetto *et al.*, 2012), Vienna (Eichinger *et al.*, 2010; Eichinger *et al.*, 2014) and HERDOO2 (Rodger *et al.*, 2008) scores are the three methods available to calculate the risk for RVTE after the three months coagulation therapy. These models were proposed using statistical techniques by analyzing several pre-selected factors and each one have its proper set of input variables.

Dash score calculates RVTE risk based on d-dimer level, age, sex, and use of hormonal therapy. Vienna score calculate the risk based on sex, location of the first VTE (distal/proximal VTE or pulmonary embolism), and d-dimer level. The score developed by Rodger *et al.* (2008) are only applied in women and takes age, d-dimer level, body mass index, and signs of hyperpigmentation, oedema or redness of either leg. It states that women with none or one of the criteria can safely discontinue anticoagulants. All scores can be used only if the patient had a first unprovoked VTE.

Reviews concerning the applicability of these scores can be found in literature (Kyrle; Eichinger, 2012; Ensor *et al.*, 2016). In a recent study, Ensor *et al.* (2016) extensively reviewed the advantages and disadvantages of each score. According to the authors, all three models were proposed using similar population data and the Dash model may have some biased conclusions, since patients with hormone intake at the time of event were also included. Besides, all scores suffer from developing limitations and lack of external validation.

External validation of any given mathematical model is crucial for practical uses. Marcucci *et al.* (2015) performed a validation study of the Vienna score (Eichinger *et al.*, 2010) and reported that it is able to distinguish patient's risk. However, they found that the model tends to underestimate the risk of RVTE at 12 months after stop coagulation therapy. Recently, Tritschler *et al.* (2015) aimed to validate the updated Vienna model (Eichinger *et al.*, 2014) for elderly patients. Their results showed that it was no possible to identify patients with low-risk of RVTE and suggest that a specific validation should be performed. Rodger *et al.* (2017) performed a validation of their proposed HERDOO2 rule and showed that it indeed works for these specific case.

Some patients develop VTE due the presence of transient risk factors. They can be reversible, such as: surgery, trauma, pregnancy, etc, or persistent, such as: cancer. It is widely accepted that these patients have a very low risk of RVTE in comparison with those with unprovoked VTE (Iorio *et al.*, 2010). However, RVTE in these conditions are still not well understood.

In a review study, Iorio *et al.* (2010) stated that the risk of RVTE appears to be low in patients with VTE provoked by a reversible risk factor and by surgery, and

intermediate in patients with VTE provoked by a nonsurgical risk factor. Besides, they affirm that both risks are very low when comparing to the patients with a previous unprovoked VTE. Cosmi *et al.* (2011) followed 296 patients with a previous provoked VTE in a 2-year study and during their study, 15 patients presented RVTE (a rate of 5.1 %). Of those 15, 10 presented higher D-dimer level after stop anticoagulation. For these patients, in 12 months, the authors reported a 7-11 % rate of recurrence, which is similar to the annual risk of RVTE for patients with unprovoked VTE. Therefore, this subject should be addressed more carefully.

This work aimed to develop a new model to predict RVTE in patients with both provoked and unprovoked previous VTE using different clinical, inherited acquired and molecular risk factors. ANN 1 included patient differentiation with respect to sex, age, as well as previous provoked or unprovoked VTE. In ANN 2, it was considered only those variables pointed as important by PCA. Both models were successful for RVTE prediction, since the results showed that correlation coefficients were very close to 1.

Two recent studies corroborate the use of ANNs for thrombosis prediction. Fei *et al.* (2017b) used a neural model to predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis. The authors trained a feed-forward backpropagation neural network with 72 patients and reported that an ANN with 9 hidden layers was more accurate than the logistic regression model. In another study, Fei *et al.* (2017a), trained radial basis ANNs to predict the risk for portal vein thrombosis in acute pancreatitis patients and obtained similar results. Qatawneh et al. (2017) trained several ANN structures to predict the risk of VTE in hospitalized patients. The authors used 35 input variables, such as: patients and disease characteristics, as well as genetic factors. The results showed that the ANNs have 81 % of accuracy.

Besides the accuracy, the models developed in this work also need external validation to offer generalization information. Once the ANNs are validated, they could be used for clinicians in day life practice. Validated models can also be used to evaluate the importance of each input parameter using Design of Experiment techniques. The result could be used to improve the number of inputs along with the PCA results. It is important to mention that the time when the blood samples were collected was not directly considered in

the model. According to Rezende *et al.* (2013), the values of the factors included in the proposed models do not fluctuate significantly with time. Thus, its influence in the dataset could be despised. Future validation studies can corroborate these findings. In summary, this work showed that the association of PCA and ANN models is efficient to model RVTE risk prediction. These equations can be developed including numerous factors and have the power to extract the interactions among them without loss of accuracy.

9 MODELOS ALTERNATIVOS

9.1 Modelo III

O ajuste referente ao Modelo III tinha por objetivo obter uma RNA que incluísse, como variável de entrada, outra informação sobre a primeira trombose do paciente: se ela era espontânea ou provocada. Conforme mencionado na Seção 7.4.6, essa informação é relevante uma vez que a literatura e os resultados deste trabalho mostraram que pacientes com trombose provocada também são passíveis de apresentarem elevados índices de TR.

O Modelo III consistiu em uma RNA com 19 variáveis de entrada e uma variável de saída. Assim como para o Modelo II, o método de Powell-Beale foi o algoritmo usado para treinar as diferentes estruturas com uma, duas e três camadas ocultas que foram testadas. Na Tabela 9.1, são apresentados a F_{OBJ} e o coeficiente de correlação para as etapas de treinamento, validação e teste, assim como o número de parâmetros para cada estrutura.

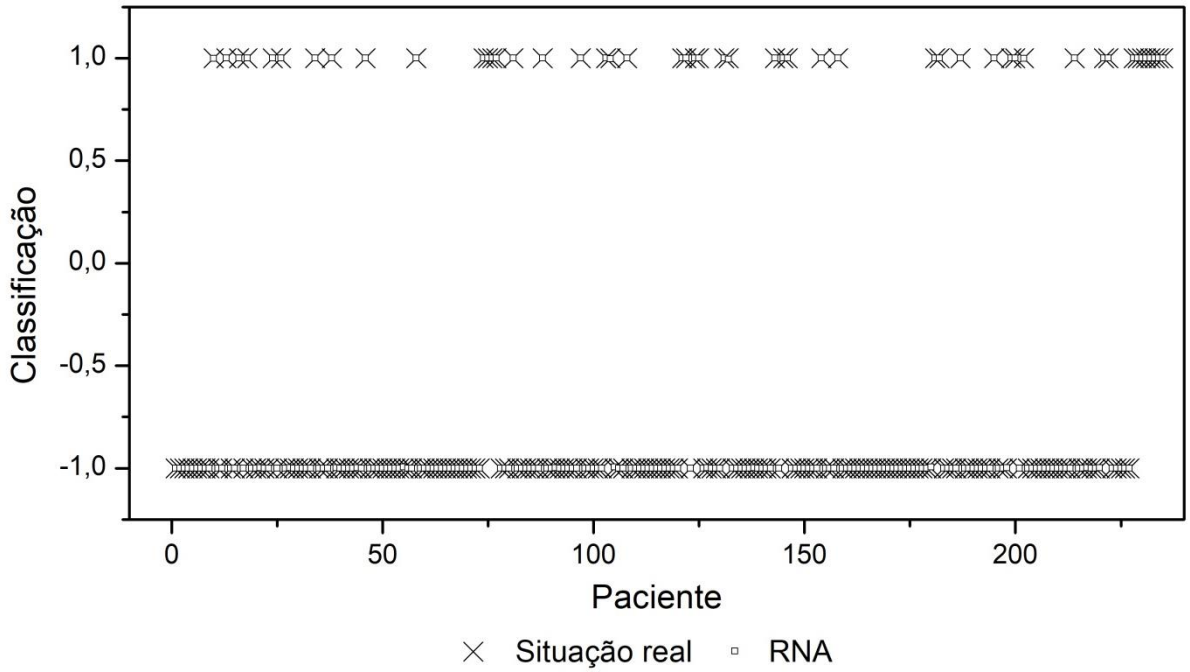
Tabela 9.1 – Resultados obtidos para as RNAs treinadas no ajuste do Modelo III.

Estrutura	F_{OBJ} Treinamento	F_{OBJ} Validação	r^2 Treinamento	r^2 Validação	r^2 Teste	Número de parâmetros
19-10-1	0,561	0,414	0,355	0,440	0,334	211
19-30-1	0,524	0,462	0,379	0,613	0,426	631
19-35-1	0,574	0,678	0,440	0,025	0,326	736
19-5-5-1	0,102	0,195	0,915	0,882	0,902	136
19-10-5-1	0,535	0,534	0,406	0,558	0,576	261
19-15-5-1	0,069	0,557	0,941	0,597	0,260	386
19-20-5-1	**$2{,}11.10^{-6}$**	**$8{,}00.10^{-8}$**	**0,999**	**0,999**	**0,999**	**511**
19-30-5-1	0,640	0,742	0,166	0,360	0,165	761
19-5-10-1	0,367	0,338	0,617	0,593	0,351	171
19-10-10-1	0,528	0,333	0,517	0,589	0,296	321
19-15-10-1	0,514	0,365	0,491	0,777	0,550	471
19-20-10-1	0,624	0,612	0,268	0,233	0,132	621
19-30-10-1	0,186	0,019	0,831	0,989	0,933	921
19-5-5-5-1	0,587	0,457	0,324	0,552	0,451	166

Os resultados apresentados na Tabela 9.1 mostram que os valores de F_{OBJ}, em geral são elevados e não seguem uma tendência de queda conforme se aumenta o número de parâmetros. A estrutura 19-20-5-1 foi a que apresentou melhor eficiência em aprender sobre o

fenômeno, assim como para prever novos casos. Como é mostrado na Tabela 9.1, essa foi a estrutura com menor valor de F_{OBJ} e maiores coeficientes de correlação em todas as etapas. Na Figura 9.1 esse resultado pode ser confirmado uma vez que a classificação real de cada paciente está sobreposta ao calculado pela RNA.

Figura 9.1 – Comparação entre a predição do Modelo III (RNA 19-20-5-1) e a classificação do paciente. Pontos das etapas de treinamento, validação e teste estão apresentados.



9.2 Modelo IV

O ajuste referente ao Modelo IV tinha por objetivo obter uma RNA que incluísse, como variável de entrada, a informação primeira trombose espontânea ou provocada, mas que excluísse os níveis de proteína C, S e antitrombina. Conforme mostrado na Seção 6.4, há a intensão de não se executar os exames de determinação dos níveis desses 3 fatores, devido a custos que podem chegar a 300 reais por paciente.

O Modelo IV consistiu em uma RNA com 16 variáveis de entrada e uma variável de saída. Assim como para os Modelos II e III, o método de Powell-Beale foi o algoritmo

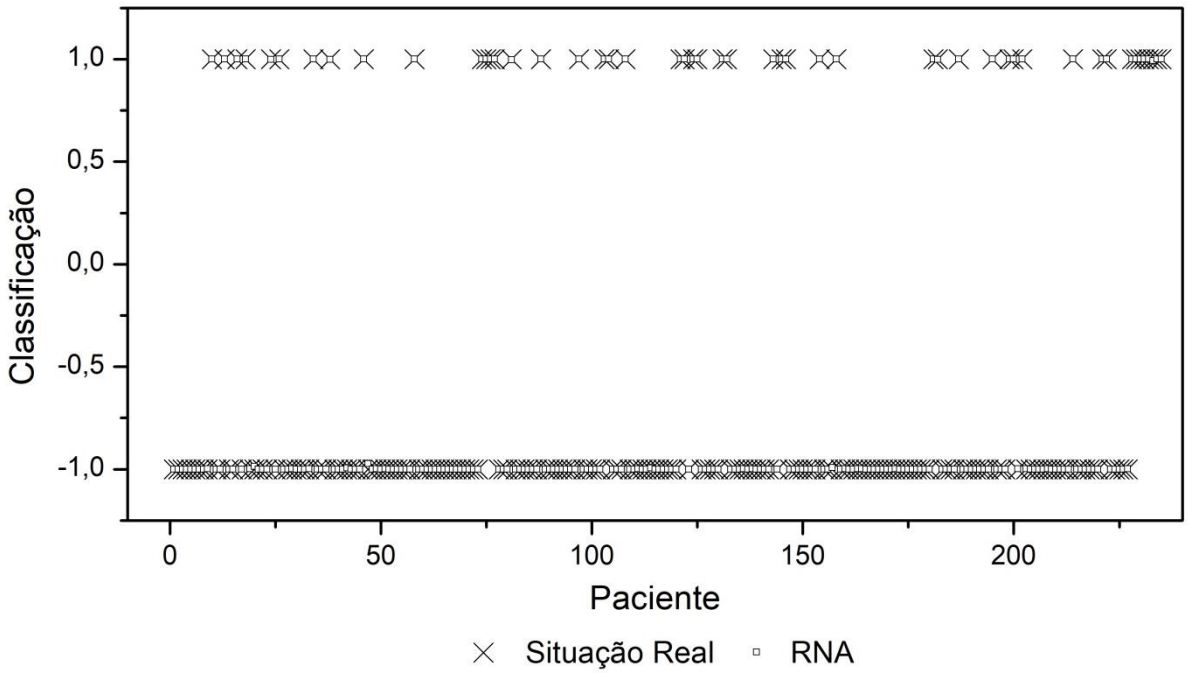
usado para treinar as diferentes estruturas com uma, duas e três camadas ocultas que foram testadas. Na Tabela 9.2 são apresentados a F_{OBJ} e o coeficiente de correlação para as etapas de treinamento, validação e teste, assim como o número de parâmetros para cada estrutura.

Tabela 9.2 – Resultados obtidos para as RNAs treinadas no ajuste do Modelo IV.

Estrutura	**F_{OBJ} Treinamento**	**F_{OBJ} Validação**	**r^2 Treinamento**	**r^2 Validação**	**r^2 Teste**	**Número de parâmetros**
16-10-1	0,301	0,349	0,774	0,646	0,370	181
16-30-1	0,598	0,632	0,260	0,423	0,403	541
16-35-1	0,465	0,497	0,538	0,558	0,587	631
16-5-5-1	0,485	0,559	0,436	0,563	0,604	121
16-5-10-1	0,398	0,605	0,657	0,403	0,661	156
16-10-5-1	0,508	0,513	0,478	0,318	0,680	231
16-15-5-1	0,198	0,114	0,843	0,918	0,614	341
16-20-5-1	0,408	0,152	0,665	0,833	0,717	451
16-30-5-1	0,540	0,235	0,450	0,740	0,431	671
16-10-10-1	0,491	0,921	0,475	0,108	0,274	291
16-15-10-1	0,073	0,228	0,941	0,849	0,999	426
16-20-10-1	0,439	0,580	0,503	0,428	0,546	561
16-30-10-1	**$6{,}15.10^{-7}$**	**$1{,}19.10^{-6}$**	**0,999**	**0,999**	**0,999**	**831**
16-5-5-5-1	0,328	0,182	0,730	0,851	0,666	151
16-10-5-5-1	0,271	0,015	0,795	0,995	0,905	261

Os resultados apresentados na Tabela 9.2 mostram que os valores de F_{OBJ}, em geral, são elevados e praticamente se mantêm constantes conforme se aumenta o número de parâmetros. Tais valores passam a diminuir para estruturas com duas camadas ocultas e 10 neurônios na segunda camada oculta. A estrutura 16-30-10-1 foi a que apresentou melhor eficiência em aprender sobre o fenômeno, assim como para prever novos casos. Como é mostrado na Tabela 9.2, essa foi a estrutura com menor valor de F_{OBJ} e maiores coeficientes de correlação em todas as etapas. Na Figura 9.2 esse resultado pode ser confirmado uma vez que a classificação real de cada paciente está sobreposta ao calculado pela RNA.

Figura 9.2 – Comparação entre a predição do Modelo IV (RNA 16-30-10-1) e a classificação do paciente. Pontos das etapas de treinamento, validação e teste estão apresentados.



9.3 Discussão

A procura por modelos matemáticos para predizer a TR com eficiência ainda é um grande desafio. Por um lado, modelos com mais fatores de entrada tem maior capacidade de generalização para diferentes casos. Por outro, existe uma desvantagem inerente à necessidade de se coletar mais parâmetros via exames clínicos – o que pode ser oneroso e ter alto custo.

Uma solução para isso é a adoção de modelos simplificados, que levem em consideração menos fatores de entrada. Essa prática pode levar a maior rapidez no resultado, bem como economizar recursos financeiros e matéria-prima. Exemplos de modelos simplificados são os métodos de *scores* propostos na literatura. Tais modelos consideram de 3 a 4 fatores de entrada. Porém uma desvantagem é que tais modelos podem falhar na previsão em certos casos, como por exemplo, quando o paciente tem idade avançada ou está no grupo de baixo risco de TR.

Este trabalho mostrou que as RNAs podem ser uma alternativa viável quando se visa obter modelos simplificados, com grande capacidade de generalização e aplicáveis nas mais diversas situações. Quatro diferentes modelos foram propostos e cada um leva em consideração diferentes fatores como entrada, desde um conjunto mais completo (Modelo I) até um conjunto reduzido em que é necessário apenas um hemograma do paciente para a predição (Modelo IV)

Para se aplicar a RNA obtida no Modelo IV, basta se conhecer as características do paciente, do primeiro episódio trombótico e o hemograma o paciente. Com esses dados, os resultados mostraram que tal equação pode prever com acurácia quais pacientes terão, ou não, um novo episódio trombótico.

Vale ressaltar que é necessário que os valores dos parâmetros de entrada, para todos os modelos propostos neste trabalho, estejam dentro do intervalo de valores em que as RNAs foram treinadas. Uma predição utilizando um fator, cujo valor esteja fora do intervalo de treinamento levará a uma predição que pode ser falsa. Possivelmente, um valor numérico bastante diferente em relação ao que se espera. Na Tabela 9.3 estão apresentados os valores possíveis para todos os fatores de entrada utilizados neste trabalho.

Tabela 9.3 – Intervalo de valores permitidos para os fatores de entrada das RNAs.

Fator	Intervalo de aplicação ou valores permitidos
Sexo(Masculino/Feminino)	1; -1
Idade (anos)	[18;82]
Embolia Pulmonar (Sim/Não)	1; -1
Lado da trombose no membro inferior (Direito/Esquerdo)	1; -1
Localização da trombose no membro inferior (Distal/Proximal)	1; -1
Provocada ou espontânea	1; -1
Tempo de anticoagulação (meses)	[1;74]
Uso de anticoagulante (Sim/Não)	1; -1
D-dímero ($ng.mL^{-1}$)	[60,65;8702,30]
FV Leiden (Heterozigoto/Não)	1; -1
Mutação G20210A gene da protrombina (Heterozigoto/Não)	1; -1
Factor VIII ($IU.dL^{-1}$)	[0,1;391,4]
Atividade de PC ($mg.dL^{-1}$)	[25;173]
Atividade de PS ($mg.dL^{-1}$)	[25,162,4]
Atividade de AT ($mg.dL^{-1}$)	[30,156,4]
Síndrome do anticorpo fosfolipídeo (Sim/Não)	1; -1
Índice de massa corporal (IMC)	[0;55,26]
Tabagismo (Sim/Não)	1; -1
Terapia hormonal (para mulher) (Sim/Não)	1; -1
Leucócitos (μL^{-1})	[3,34;37,02]
Hemácias (μL^{-1})	[1,25;7,34]
Hemoglobina ($mg.dL^{-1}$)	[12,4;17,5]
Hematócrito (%)	[10,05;53,8]
Distribuição de tamanho dos eritrócitos (%)	[10,05;23]
Plaquetas (μL^{-1})	[48;535]
Volume médio das plaquetas (fL)	[6;13,3]
Colesterol total ($mg.dL^{-1}$)	[9;326]
Lipoproteína de alta densidade (HDL) ($mg.dL^{-1}$)	[24;110]
Lipoproteína de baixa densidade (LDL) ($mg.dL^{-1}$)	[40;191]
Triglicérides ($mg.dL^{-1}$)	[0,05;1522]
Glicose ($mg.dL^{-1}$)	[0,34;391]
Creatinina ($mg.dL^{-1}$)	[0,02;2,8]
Proteína C reativa ($mg.dL^{-1}$)	[0,02;255]
Hipertensão arterial (Sim/Não)	1; -1
Diabetes (Sim/Não)	1; -1
Dislipidemia (Sim/Não)	1; -1
Insuficiência Renal (Sim/Não)	1; -1
Cancer (Sim/Não)	1; -1
Trombo residual (Sim/Não)	1; -1

10 CONCLUSÕES

A TR é uma complicação que acomete até 25 % dos indivíduos nos primeiros 5 anos após o primeiro episódio trombótico. Assim, os principais objetivos deste trabalho foram: i) determinar os principais fatores preditivos da TR e, ii) obter modelos matemáticos para predizer a TR empregando RNAs. Tais modelos poderão ser utilizados futuramente na prevenção dessa complicação, auxiliando médicos na tomada de decisão a respeito da continuidade ou não do tratamento anticoagulante.

10.1 Primeiro Manuscrito

O primeiro manuscrito teve como objetivo a determinação dos fatores preditivos mais importantes na predição da TR. Para isso, a técnica de ACP foi empregada em um conjunto de dados contendo 39 fatores pré-selecionados. Como resultado, a técnica apontou que 13 componentes principais eram o suficiente para descrever o problema e, através da análise do coeficiente de correlação entre os fatores e as componentes principais, verificou-se que 18 são os mais importantes para a predição da TR. Esta etapa do trabalho mostrou que a aplicação de uma técnica estatística não muito comum na Medicina pode auxiliar na determinação de fatores preditivos para as mais diversas enfermidades e suas complicações.

10.2 Segundo Manuscrito

O Segundo manuscrito descreveu o desenvolvimento de dois modelos neurais para a predição da TR. Para isso, diferentes estruturas de RNAs foram treinadas com 3 algoritmos de otimização, considerando dois conjuntos de dados de entrada: i) os 39 fatores pré-selecionados e, ii) os 18 fatores apontados como principais pela ACP. Os resultados mostraram que as RNAs são capazes de predizer quais pacientes irão desenvolver TR para ambos os casos. As RNAs apresentaram excelente correlação e precisão em relação à classificação dos pacientes. Esta etapa do trabalho mostrou que as RNAs são uma ferramenta poderosa que pode auxiliar os médicos na tomada a decisão em relação à continuidade do tratamento anticoagulante.

10.3 Modelos Alternativos

O Capítulo 9 descreveu o desenvolvimento de dois modelos neurais alternativos para a predição da TR. Tais modelos se diferem dos anteriores em seu conjunto de variáveis de entrada. No modelo III foram consideradas 19 variáveis de entrada, sendo elas as mesmas do Modelo II, incluindo-se o fator TVP espontânea/provocada, devido a sua importância no cotidiano. No modelo IV foram consideradas 16 variáveis de entrada, sendo elas as mesmas do Modelo III, exceto proteína C, S e antitrombina. Os resultados mostraram que as RNAs são capazes de predizer quais pacientes irão desenvolver TR para ambos os casos, apresentando excelente correlação e precisão em relação à classificação dos pacientes. Esta etapa do trabalho confirmou que as RNAs são uma ferramenta que podem se adaptar a diferentes situações sem perda de eficiência.

10.4 Limitações Encontradas e Considerações Finais

As principais limitações deste projeto dizem respeito à coleta de dados. Alguns pacientes tiveram que ser excluídos devido à falta de informações a respeito dos exames, visto que alguns dos parâmetros não faziam parte da rotina de avaliação. Além disso, outra dificuldade encontrada foi decorrente do fato da necessidade de acessar diferentes bancos de dados na coleta de informações, por não serem integrados. Um fator não considerado nos modelos é o tempo entre o fim da anticoagulação e os exames coletados, conforme citado nas Seções 7.4.6 e 8.4.

Este trabalho mostrou que a associação de técnicas estatísticas e de inteligência artificial, comumente usadas na Engenharia, gerou uma ferramenta importante para a predição e prevenção da TR numa população brasileira, que pode ser validada em outras populações: diversas RNAs capazes de classificar que pacientes terão TR e quais não. Além disso, mostraram que a Análise de Componentes Principais pode ajudar na identificação dos principais fatores da TR, gerando modelos neurais com menos entradas, que são de baixo custo, e que apresentam a mesma eficiência em relação ao modelo com mais entradas.

Após a devida validação dos modelos, as RNAs poderão ser empregadas na predição de TR. Além disso, vale ressaltar que as RNAs podem ser atualizadas conforme

novos casos são incluídos no banco de dados. Isso é feito retreinando-se as melhores estruturas obtidas, incluindo esses novos casos no conjunto de dados de treinamento. Com isso, modelos cada vez mais precisos e com maior capacidade de generalização podem ser obtidos.

Por fim, neste trabalho foram propostos 4 diferentes modelos neurais, capazes de classificar eficientemente pacientes propensos a TR com base em diferentes conjuntos de parâmetros clínicos. A escolha de qual modelo se aplicar pode ser em função da possibilidade de coleta dos fatores necessários. Conforme os resultados apresentados, a eficiência de todos os modelos foi atestada nas etapas de validação e teste. Porém, recomenda-se uma validação externa para confirmação dos resultados.

10.5 Sugestões para Trabalhos Futuros

Este trabalho foi o início de uma parceria entre a Faculdade de Engenharia Química e a Faculdade de Ciências Médicas, da Unicamp, no campo da inteligência artificial. A partir dos resultados encontrados neste trabalho, surgiram diversos desdobramentos que poderão ser abordados em trabalhos futuros e seguem abaixo:

1. Realizar um estudo prospectivo com um novo grupo de pacientes de modo a se incluir o tempo entre o fim do tratamento de anticoagulação e a coleta dos exames, avaliando-se o efeito dessa variável no modelo;
2. Validar os modelos desenvolvidos neste trabalho em diferentes grupos de pacientes acompanhados em outros serviços;
3. Gerar um software para o uso cotidiano dos modelos obtidos;
4. Realizar um estudo de análise de sensibilidade dos fatores empregando a técnica estatística de Plackett-Burman;
5. Desenvolver um modelo neuro-*fuzzy* para a predição da TR;
6. Desenvolver um modelo neural, e/ou de neuro-*fuzzy*, para prever qual o tempo de tratamento para pacientes com baixo risco de TR;
7. Desenvolver um modelo neural, e/ou de neuro-*fuzzy*, para prever o risco de sangramento;

8. Desenvolver um modelo neural, e/ou de neuro-*fuzzy*, para prever a influência dos diferentes anticoagulantes na prevenção da TR;
9. Desenvolver um modelo neural, e/ou de neuro-*fuzzy*, para prever a TR em pacientes com câncer;
10. Desenvolver um modelo neural, e/ou de neuro-*fuzzy*, para prever a TR em pacientes com SAF.

11 REFERÊNCIAS

ABEDI, V.; GOYAL, N.; TSIVGOULIS, G.; HOSSEINICHIMEH, N.; HONTECILLAS, R.; BASSAGANYA-RIERA, J.; ELIJOVICH, L.; METTER, J. E.; ALEXANDROV, A. W.; LIEBESKIND, D. S.; ALEXANDROV, A. V.; ZAND, R. Novel Screening Tool for Stroke Using Artificial Neural Network. **Stroke**, v. 48, n. 6, p. 1678-1681, 2017.

AGGARWAL, A.; PURI, K.; LIANGPUNSAKUL, S. Deep vein thrombosis and pulmonary embolism in cirrhotic patients: Systematic review. **World Journal of Gastroenterology : WJG**, v. 20, n. 19, p. 5737-5745, 2014.

AGNELLI, G.; BULLER, H. R.; COHEN, A.; CURTO, M.; GALLUS, A. S.; JOHNSON, M.; PORCARI, A.; RASKOB, G. E.; WEITZ, J. I. Apixaban for Extended Treatment of Venous Thromboembolism. **New England Journal of Medicine**, v. 368, n. 8, p. 699-708, 2013.

ALT, E.; BANYAI, S.; BANYAI, M.; KOPPENSTEINER, R. Blood rheology in deep venous thrombosis—relation to persistent and transient risk factors. **Thrombosis Research**, v. 107, n. 3, p. 101-107, 2002.

ANTHONY LIZARRAGA, W.; DALIA, S.; REINERT, S. E.; SCHIFFMAN, F. J. Venous thrombosis in patients with chronic liver disease. **Blood Coagulation & Fibrinolysis**, v. 21, n. 5, p. 431-435, 2010.

BABU, M.; RAMARAJ, N.; RAJAGOPALAN, S. Heart diseases data classification using group search optimisation with artificial neural network approach. **International Journal of Business Intelligence and Data Mining**, v. 12, n. 3, p. 257-273, 2017.

BADNJEVIĆ, A.; GURBETA, L.; CIFREK, M.; MARJANOVIC, D. **Classification of asthma using artificial neural network**. 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). May 30 2016-June 3 2016, 2016.

BAGOT, C. N.; ARYA, R. Virchow and his triad: a question of attribution. **British Journal of Haematology**, v. 143, n. 2, p. 180-190, 2008.

BARILLARI, G.; LONDERO, A. P.; BRENNER, B.; NAUFFAL, D.; MUÑOZ-TORRERO, J. F. S.; DEL MOLINO, F.; MOUSTAFA, F.; MADRIDANO, O.; MARTÍN-MARTOS, F.; MONREAL, M. Recurrence of venous thromboembolism in patients with recent gestational deep vein thrombosis or pulmonary embolism: Findings from the RIETE Registry. **European Journal of Internal Medicine**, v. 32, n. Supplement C, p. 53-59, 2016.

BEALE, E. A derivation of conjugate gradients. **Numerical methods for nonlinear optimization**, p. 39-43, 1972.

BELAJ, K.; HACKL, G.; RIEF, P.; ELLER, P.; BRODMANN, M.; GARY, T. Changes in Lipid Metabolism and Extension of Venous Thromboembolism. **Annals of Nutrition and Metabolism**, v. 64, n. 2, p. 122-126, 2014.

BEYTH, R. J.; COHEN, A. M.; LANDEFELD, C. Long-term outcomes of deep-vein thrombosis. **Archives of Internal Medicine**, v. 155, n. 10, p. 1031-1037, 1995.

BEZERRA, S. **Reservoir Computing com Hierarquia para Previsão de Vazões Médias Diárias**. Dissertação (Mestrado). Universidade Federal de Pernambuco, 2016.

BJØRI, E.; JOHNSEN, H. S.; HANSEN, J. B.; BRÆKKAN, S. K. D-dimer at venous thrombosis diagnosis is associated with risk of recurrence. **Journal of Thrombosis and Haemostasis**, v. 15, n. 5, p. 917-924, 2017.

BOUTITIE, F.; PINEDE, L.; SCHULMAN, S.; AGNELLI, G.; RASKOB, G.; JULIAN, J.; HIRSH, J.; KEARON, C. Influence of preceding length of anticoagulant treatment and initial presentation of venous thromboembolism on risk of recurrence after stopping treatment: analysis of individual participants' data from seven trials. **BMJ**, v. 342, 2011.

BRÆKKAN, S. K.; MATHIESEN, E. B.; NJØLSTAD, I.; WILSGAARD, T.; HANSEN, J.-B. Hematocrit and risk of venous thromboembolism in a general population. The Tromsø study. **Haematologica**, v. 95, n. 2, p. 270-275, 2010.

BRAGA, A. P.; CARVALHO, A. P. L.; LUDERMIR, T. B. **Redes Neurais Artificiais: Teoria e Aplicações**. 2ª ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos, 2007.

BUCCIARELLI, P.; MAINO, A.; FELICETTA, I.; ABBATTISTA, M.; PASSAMONTI, S. M.; ARTONI, A.; MARTINELLI, I. Association between red cell distribution width and risk of venous thromboembolism. **Thrombosis Research**, v. 136, n. 3, p. 590-594, 2015.

BYRNES, J. R.; WOLBERG, A. S. Red blood cells in thrombosis. **Blood**, v. 130, n. 16, p. 1795-1799, 2017.

CAROBBIO, A.; FINAZZI, G.; GUERINI, V.; SPINELLI, O.; DELAINI, F.; MARCHIOLI, R.; BORRELLI, G.; RAMBALDI, A.; BARBUI, T. Leukocytosis is a risk factor for thrombosis in essential thrombocythemia: interaction with treatment, standard risk factors, and Jak2 mutation status. **Blood**, v. 109, n. 6, p. 2310-2313, 2007.

CARRIER, M.; RODGER, M. A.; WELLS, P. S.; RIGHINI, M.; LE GAL, G. Residual vein obstruction to predict the risk of recurrent venous thromboembolism in patients with deep vein thrombosis: a systematic review and meta-analysis. **Journal of Thrombosis and Haemostasis**, v. 9, n. 6, p. 1119-1125, 2011.

CHIN, T. L.; MOORE, E. E.; MOORE, H. B.; GONZALEZ, E.; CHAPMAN, M. P.; STRINGHAM, J. R.; RAMOS, C. R.; BANERJEE, A.; SAUAIA, A. A principal component analysis of postinjury viscoelastic assays: Clotting factor depletion versus fibrinolysis. **Surgery**, v. 156, n. 3, p. 570-577, 2014.

CHRISTIANSEN, S. C.; LIJFERING, W. M.; HELMERHORST, F. M.; ROSENDAAL, F. R.; CANNEGIETER, S. C. Sex difference in risk of recurrent venous thrombosis and the risk profile for a second event. **Journal of Thrombosis and Haemostasis**, v. 8, n. 10, p. 2159-2168, 2010.

CHUNG, W.-S.; LIN, C.-L.; KAO, C.-H. Diabetes increases the risk of deep-vein thrombosis and pulmonary embolism. **Thrombosis and haemostasis**, v. 114, n. 4, p. 812-818, 2015.

COMARMOND, C.; JEGO, P.; VEYSSIER-BELOT, C.; MARIE, I.; MEKINIAN, A.; ELMALEH-SACHS, A.; LEROUX, G.; SAADOUN, D.; OZIOL, E.; FRAISSE, T.; HYVERNAT, H.; THIERCEIN-LEGRAND, M. F.; SARROT-REYNAULD, F.; FERREIRA-MALDENT, N.; DE MENTHON, M.; GOUJARD, C.; KHAU, D.; NGUEN, Y.; MONNIER, S.; MICHON, A.; CASTEL, B.; DECAUX, O.; PIETTE, J. C.; CACOUB, P. Cessation of oral anticoagulants in antiphospholipid syndrome. **Lupus**, v. 26, n. 12, p. 1291-1296, 2017.

COSMI, B.; LEGNANI, C.; CINI, M.; GUAZZALOCA, G.; PALARETI, G. D-dimer and residual vein obstruction as risk factors for recurrence during and after anticoagulation withdrawal in patients with a first episode of provoked deep-vein thrombosis. **Thrombosis and haemostasis**, v. 105, n. 5, p. 837, 2011.

CUSHMAN, M.; O'MEARA, E. S.; HECKBERT, S. R.; ZAKAI, N. A.; ROSAMOND, W.; FOLSOM, A. R. Body size measures, hemostatic and inflammatory markers and risk of venous thrombosis: The Longitudinal Investigation of Thromboembolism Etiology. **Thrombosis Research**, v. 144, n. Supplement C, p. 127-132, 2016.

DAVIE, E. W.; RATNOFF, O. D. Waterfall Sequence for Intrinsic Blood Clotting. **Science**, v. 145, n. 3638, p. 1310-1312, 1964.

DE STEFANO, V.; SIMIONI, P.; ROSSI, E.; TORMENE, D.; ZA, T.; PAGNAN, A.; LEONE, G. The risk of recurrent venous thromboembolism in patients with inherited deficiency of natural anticoagulants antithrombin, protein C and protein S. **Haematologica**, v. 91, n. 5, p. 695, 2006.

DE STEFANO, V.; ZA, T.; ROSSI, E.; VANNUCCHI, A. M.; RUGGERI, M.; ELLI, E.; MICÒ, C.; TIEGHI, A.; CACCIOLA, R. R.; SANTORO, C.; GERLI, G.; VIANELLI, N.; GUGLIELMELLI, P.; PIERI, L.; SCOGNAMIGLIO, F.; RODEGHIERO, F.; POGLIANI, E. M.; FINAZZI, G.; GUGLIOTTA, L.; MARCHIOLI, R.; LEONE, G.; BARBUI, T. Recurrent thrombosis in patients with polycythemia vera and essential thrombocythemia: incidence, risk factors, and effect of treatments. **Haematologica**, v. 93, n. 3, p. 372-380, 2008.

DEGUCHI, H.; PECHENIUK, N. M.; ELIAS, D. J.; AVERELL, P. M.; GRIFFIN, J. H. High-Density Lipoprotein Deficiency and Dyslipoproteinemia Associated With Venous Thrombosis in Men. **Circulation**, v. 112, n. 6, p. 893, 2005.

DINTENFASS, L. Viscosity and Clotting of Blood in Venous Thrombosis and Coronary Occlusions. **Circulation Research**, v. 14, n. 1, p. 1, 1964.

EICHINGER, S.; HEINZE, G.; JANDECK, L. M.; KYRLE, P. A. Risk Assessment of Recurrence in Patients With Unprovoked Deep Vein Thrombosis or Pulmonary Embolism: The Vienna Prediction Model. **Circulation**, v. 121, n. 14, p. 1630-1636, 2010.

EICHINGER, S.; HEINZE, G.; KYRLE, P. A. d-Dimer Levels Over Time and the Risk of Recurrent Venous Thromboembolism: An Update of the Vienna Prediction Model. **Journal of the American Heart Association**, v. 3, n. 1, 2014.

EICHINGER, S.; HRON, G.; BIALONCZYK, C.; ET AL. Overweight, obesity, and the risk of recurrent venous thromboembolism. **Archives of Internal Medicine**, v. 168, n. 15, p. 1678-1683, 2008.

EICHINGER, S.; MINAR, E.; BIALONCZYK, C.; ET AL. D-dimer levels and risk of recurrent venous thromboembolism. **JAMA**, v. 290, n. 8, p. 1071-1074, 2003.

EICHINGER, S.; PECHENIUK, N. M.; HRON, G.; DEGUCHI, H.; SCHEMPER, M.; KYRLE, P. A.; GRIFFIN, J. H. High-Density Lipoprotein and the Risk of Recurrent Venous Thromboembolism. **Circulation**, v. 115, n. 12, p. 1609, 2007.

EISCHER, L.; TSCHOLL, V.; HEINZE, G.; TRABY, L.; KYRLE, P. A.; EICHINGER, S. Hematocrit and the Risk of Recurrent Venous Thrombosis: A Prospective Cohort Study. **PLOS ONE**, v. 7, n. 6, p. e38705, 2012.

ENSOR, J.; RILEY, R. D.; MOORE, D.; SNELL, K. I. E.; BAYLISS, S.; FITZMAURICE, D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. **BMJ Open**, v. 6, n. 5, 2016.

ESMON, C. The protein C pathway. **Critical Care Medicine**, v. 28, n. 9, p. S44-S48, 2000.

EVERETT, B. M.; GLYNN, R. J.; BURING, J. E.; RIDKER, P. M. Lipid biomarkers, hormone therapy, and the risk of venous thromboembolism in women. **Journal of thrombosis and haemostasis : JTH**, v. 7, n. 4, p. 588-596, 2009.

FAHRNI, J.; HUSMANN, M.; GRETENER, S. B.; KEO, H. H. Assessing the risk of recurrent venous thromboembolism – a practical approach. **Vascular Health and Risk Management**, v. 11, p. 451-459, 2015.

FARZAMNIA, H.; RABIEI, K.; SADEGHI, M.; ROGHANI, F. The Predictive Factors of Recurrent Deep Vein Thrombosis. **ARYA Atherosclerosis**, v. 7, n. 3, p. 123-128, 2011.

FEI, Y.; HU, J.; GAO, K.; TU, J.; LI, W.-Q.; WANG, W. Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models. **Journal of Critical Care**, v. 39, p. 115-123, 2017a.

FEI, Y.; HU, J.; LI, W. Q.; WANG, W.; ZONG, G. Q. Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis. **Journal of Thrombosis and Haemostasis**, v. 15, n. 3, p. 439-445, 2017b.

FERNANDES, C. J. C. D. S.; ALVES JÚNIOR, J. L.; GAVILANES, F.; PRADA, L. F.; MORINAGA, L. K.; SOUZA, R. New anticoagulants for the treatment of venous thromboembolism. **Jornal Brasileiro de Pneumologia**, v. 42, p. 146-154, 2016.

FERREIRA, C. N.; SOUSA, M. D. O.; DUSSE, L. M. S. A.; CARVALHO, M. D. G. O novo modelo da cascata de coagulação baseado nas superfícies celulares e suas implicações. **Revista Brasileira de Hematologia e Hemoterapia**, v. 32, p. 416-421, 2010.

FRANCO, L.; GIUSTOZZI, M.; AGNELLI, G.; BECATTINI, C. Anticoagulation in patients with isolated distal deep vein thrombosis: a meta-analysis. **Journal of Thrombosis and Haemostasis**, v. 15, n. 6, p. 1142-1154, 2017.

GALANAUD, J. P.; SEVESTRE, M. A.; GENTY, C.; KAHN, S. R.; PERNOD, G.; ROLLAND, C.; DIARD, A.; DUPAS, S.; JURUS, C.; DIAMAND, J. M.; QUERE, I.; BOSSON, J. L.; FOR THE, O.-S. I. Incidence and predictors of venous thromboembolism recurrence after a first isolated distal deep vein thrombosis. **Journal of Thrombosis and Haemostasis**, v. 12, n. 4, p. 436-443, 2014.

GANGAT, N.; STRAND, J.; LI, C.-Y.; WU, W.; PARDANANI, A.; TEFFERI, A. Leucocytosis in polycythaemia vera predicts both inferior survival and leukaemic transformation. **British Journal of Haematology**, v. 138, n. 3, p. 354-358, 2007.

GARCÍA, R. A.; ENE, G.; MIRANDA, C.; VIDAL, R.; MATA, R.; LLAMAS, S. M. Association between venous thrombosis and dyslipidemia. **Medicina clinica**, v. 143, n. 1, p. 1-5, 2014.

GHAVAMI, P.; KAPUR, K. **Prognostics and artificial neural network applications in patient healthcare**. Prognostics and Health Management (PHM), 2011 IEEE Conference on. 20-23 June 2011, 2011.

GRANT, P. J. Diabetes mellitus as a prothrombotic condition. **Journal of Internal Medicine**, v. 262, n. 2, p. 157-172, 2007.

GUYATT, G. H.; AKL, E. A.; CROWTHER, M.; GUTTERMAN, D. D.; SCHUÜNEMANN, H. J.; FOR THE AMERICAN COLLEGE OF CHEST PHYSICIANS ANTITHROMBOTIC, T.; PREVENTION OF THROMBOSIS, P. Executive Summary: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. **Chest**, v. 141, n. 2 Suppl, p. 7S-47S, 2012.

HANSSON, P.; SÖRBO, J.; ERIKSSON, H. Recurrent venous thromboembolism after deep vein thrombosis: Incidence and risk factors. **Archives of Internal Medicine**, v. 160, n. 6, p. 769-774, 2000.

HÄRDLE, W.; SIMAR, L. **Applied multivariate statistical analysis**. 2 ed. Berlin: Springer Berlin Heidelberg, 2007.

HAYKIN, S. **Neural Networks – A Comprehensive Foundation**. Delhi: Prentice Hall, 2005.

HEEB, M. J.; ROSING, J.; BAKKER, H. M.; FERNANDEZ, J. A.; TANS, G.; GRIFFIN, J. H. Protein S binds to and inhibits factor Xa. **Proceedings of the National Academy of Sciences**, v. 91, n. 7, p. 2728-2732, 1994.

HEIT, J. A. Predicting the Risk of Venous Thromboembolism Recurrence. **American journal of hematology**, v. 87, n. Suppl 1, p. S63-S67, 2012.

HEIT, J. A.; MOHR, D. N.; SILVERSTEIN, M. D.; PETTERSON, T. M.; O'FALLON, W.; MELTON, L.; III. Predictors of recurrence after deep vein thrombosis and pulmonary embolism: A population-based cohort study. **Archives of Internal Medicine**, v. 160, n. 6, p. 761-768, 2000.

HESS, K.; GRANT, P. J. Inflammation and thrombosis in diabetes. **Thrombosis and haemostasis**, v. 105, 2011.

HIRSH, J.; HOAK, J. Management of Deep Vein Thrombosis and Pulmonary Embolism. **A Statement for Healthcare Professionals From the Council on Thrombosis (in Consultation With the Council on Cardiovascular Radiology), American Heart Association**, v. 93, n. 12, p. 2212-2245, 1996.

HOFFMAN, M.; MONROE III, D. M. A cell-based model of hemostasis. **Thrombosis and haemostasis**, v. 85, n. 06, p. 958-965, 2001.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, v. 2, n. 5, p. 359-366, 1989.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of educational psychology**, v. 24, n. 6, p. 417, 1933a.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 7, p. 498-520, 1933b.

HUANG, L.; LI, J.; JIANG, Y. Association between hypertension and deep vein thrombosis after orthopedic surgery: a meta-analysis. **European Journal of Medical Research**, v. 21, p. 13, 2016.

HULL , R.; DELMORE , T.; GENTON , E.; HIRSH , J.; GENT , M.; SACKETT , D.; MCLOUGHLIN , D.; ARMSTRONG , P. Warfarin Sodium versus Low-Dose Heparin in the Long-Term Treatment of Venous Thrombosis. **New England Journal of Medicine**, v. 301, n. 16, p. 855-858, 1979.

IBOPE. Resultados pesquisa IBOPE: Trombose venose profunda e embolia pulmonar. 2010.

INVESTIGATORS, T. E. Oral Rivaroxaban for Symptomatic Venous Thromboembolism. **New England Journal of Medicine**, v. 363, n. 26, p. 2499-2510, 2010.

IORIO, A.; KEARON, C.; FILIPPUCCI, E.; ET AL. Risk of recurrence after a first episode of symptomatic venous thromboembolism provoked by a transient risk factor: A systematic review. **Archives of Internal Medicine**, v. 170, n. 19, p. 1710-1716, 2010.

KEARON, C.; AKL, E. A.; ORNELAS, J.; BLAIVAS, A.; JIMENEZ, D.; BOUNAMEAUX, H.; HUISMAN, M.; KING, C. S.; MORRIS, T. A.; SOOD, N.; STEVENS, S. M.; VINTCH, J. R. E.; WELLS, P.; WOLLER, S. C.; MOORES, L. Antithrombotic Therapy for VTE Disease. **CHEST**, v. 149, n. 2, p. 315-352, 2016.

KEARON, C.; KAHN, S. R.; AGNELLI, G.; GOLDHABER, S.; RASKOB, G. E.; COMEROTA, A. J. Antithrombotic Therapy for Venous Thromboembolic Disease. **CHEST**, v. 133, n. 6, p. 454S-545S, 2008.

KIM, H. K.; KIM, J. E.; PARK, S. H.; KIM, Y. I.; NAM-GOONG, I. S.; KIM, E. S. High coagulation factor levels and low protein C levels contribute to enhanced thrombin generation in patients with diabetes who do not have macrovascular complications. **Journal of Diabetes and its Complications**, v. 28, n. 3, p. 365-369, 2014.

KLOK, F.; NIEMANN, C.; DELLAS, C.; HASENFUß, G.; KONSTANTINIDES, S.; LANKEIT, M. Performance of five different bleeding-prediction scores in patients with acute pulmonary embolism. **Journal of thrombosis and thrombolysis**, v. 41, n. 2, p. 312-320, 2016.

KOJURI, J.; BOOSTANI, R.; DEHGHANI, P.; NOWROOZIPOUR, F.; SAKI, N. Prediction of Acute Myocardial infarction with Artificial Neural Networks in Patients with Nondiagnostic Electrocardiogram. **Journal of Cardiovascular Disease Research Vol**, v. 6, n. 2, p. 51, 2015.

KUMAR, M.; SHARMA, A.; AGARWAL, S. **Clinical decision support system for diabetes disease diagnosis using optimized neural network**. Engineering and Systems (SCES), 2014 Students Conference on. 28-30 May 2014, 2014.

KUMARI, N.; SUNITA, S. Comparison of ANNs, Fuzzy Logic and Neuro-Fuzzy Integrated Approach for Diagnosis of Coronary Heart Disease: A Survey. 2013.

KUTCHER, M. E.; FERGUSON, A. R.; COHEN, M. J. A principal component analysis of coagulation after trauma. **The journal of trauma and acute care surgery**, v. 74, n. 5, p. 1223-1230, 2013.

KYRLE, P. A. How I treat recurrent deep-vein thrombosis. **Blood**, v. 127, n. 6, p. 696-702, 2016.

KYRLE, P. A.; EICHINGER, S. Is Virchow's triad complete? **Blood**, v. 114, n. 6, p. 1138-1139, 2009.

KYRLE, P. A.; EICHINGER, S. Clinical scores to predict recurrence risk of venous thromboembolism. **Thrombosis and haemostasis**, v. 108, n. 6, p. 1061, 2012.

KYRLE , P. A.; MINAR , E.; BIALONCZYK , C.; HIRSCHL , M.; WELTERMANN , A.; EICHINGER , S. The Risk of Recurrent Venous Thromboembolism in Men and Women. **New England Journal of Medicine**, v. 350, n. 25, p. 2558-2563, 2004.

LACROIX, D. A reduced equation mathematical model for blood coagulation and fibrinolysis in quiescent plasma. **The International Journal of Structural Changes in Solids**, v. 4, p. 23-35, 2012.

LEIDERMAN, K.; FOGELSON, A. An overview of mathematical modeling of thrombus formation under flow. **Thrombosis Research**, v. 133, n. Supplement 1, p. S12-S14, 2014.

LELA, M.; SALAH, A.-M.; PEARCE, G.; TAMAR, G.; JAVAKHISHVILI, I. Blood clotting prediction model using Artificial Neural Networks and Sensor Networks. **GESJ: Computer Science and Telecommunications**, n. 3, p. 43, 2014.

LETTERIE, G. S.; BORSETH, J. An artificial neural network (ANN) as a clinical decision making tool during ovarian stimulation and in vitro fertiliztion (IVF). **Fertility and Sterility**, v. 102, n. 3, p. e112, 2014.

LIJFERING, W. M.; ROSENDAAL, F. R.; CANNEGIETER, S. C. Risk factors for venous thrombosis – current understanding from an epidemiological point of view. **British Journal of Haematology**, v. 149, n. 6, p. 824-833, 2010.

LIJFERING, W. M.; VEEGER, N. J. G. M.; MIDDELDORP, S.; HAMULYÁK, K.; PRINS, M. H.; BÜLLER, H. R.; VAN DER MEER, J. A lower risk of recurrent venous thrombosis in women compared with men is explained by sex-specific risk factors at time of first venous thrombosis in thrombophilic families. **Blood**, v. 114, n. 10, p. 2031-2036, 2009.

LIPPI, G.; FAVALORO, E. J.; CERVELLIN, G. **A review of the value of D-dimer testing for prediction of recurrent venous thromboembolism with increasing age**. Seminars in thrombosis and hemostasis: Thieme Medical Publishers, 2014.

LITVINOV, R. I.; WEISEL, J. W. Role of red blood cells in haemostasis and thrombosis. **ISBT Science Series**, v. 12, n. 1, p. 176-183, 2017.

LOPRESTI, R.; FERRARA, F.; CANINO, B.; MONTANA, M.; CAIMI, G. Deep venous thrombosis: leukocyte rheology at baseline and after in vitro activation. **Pathophysiology of Haemostasis and Thrombosis**, v. 30, n. 4, p. 168-173, 2000.

MACFARLANE, R. G. An Enzyme Cascade in the Blood Clotting Mechanism, and its Function as a Biochemical Amplifier. **Nature**, v. 202, n. 4931, p. 498-499, 1964.

MARCHIOLI , R.; FINAZZI , G.; SPECCHIA , G.; CACCIOLA , R.; CAVAZZINA , R.; CILLONI , D.; DE STEFANO , V.; ELLI , E.; IURLO , A.; LATAGLIATA , R.; LUNGHI , F.; LUNGHI , M.; MARFISI , R. M.; MUSTO , P.; MASCIULLI , A.; MUSOLINO , C.; CASCAVILLA , N.; QUARTA , G.; RANDI , M. L.; RAPEZZI , D.; RUGGERI , M.; RUMI , E.; SCORTECHINI , A. R.; SANTINI , S.; SCARANO , M.; SIRAGUSA , S.; SPADEA ,

A.; TIEGHI , A.; ANGELUCCI , E.; VISANI , G.; VANNUCCHI , A. M.; BARBUI , T. Cardiovascular Events and Intensity of Treatment in Polycythemia Vera. **New England Journal of Medicine**, v. 368, n. 1, p. 22-33, 2013.

MARCUCCI, M.; IORIO, A.; DOUKETIS, J. D.; EICHINGER, S.; TOSETTO, A.; BAGLIN, T.; CUSHMAN, M.; PALARETI, G.; POLI, D.; TAIT, R. C.; KYRLE, P. A. Risk of recurrence after a first unprovoked venous thromboembolism: external validation of the Vienna Prediction Model with pooled individual patient data. **Journal of Thrombosis and Haemostasis**, v. 13, n. 5, p. 775-781, 2015.

MARQUARDT, D. W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. **Journal of the Society for Industrial and Applied Mathematics**, v. 11, n. 2, p. 431-441, 1963.

MARTINS, T. D.; ROMANO, A. V. C.; ANNICHINO-BIZZACCHI, J. M.; FILHO, R. M. Principal Component Analysis on Recurrent Venous Thrombosis. **Manuscript submitted for publication**, 2017.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, n. 4, p. 115-133, 1943.

MCRAE, S.; TRAN, H.; SCHULMAN, S.; GINSBERG, J.; KEARON, C. Effect of patient's sex on risk of recurrent venous thromboembolism: a meta-analysis. **The Lancet**, v. 368, n. 9533, p. 371-378, 2006.

MÉAN, M.; LIMACHER, A.; STALDER, O.; ANGELILLO-SCHERRER, A.; ALBERIO, L.; FONTANA, P.; BEER, H.-J.; RODONDI, N.; LÄMMLE, B.; AUJESKY, D. Do Factor V Leiden and Prothrombin G20210A Mutations Predict Recurrent Venous Thromboembolism in Older Patients? **The American Journal of Medicine**, v. 130, n. 10, p. 1220.e17-1220.e22, 2017.

MEDINA, G.; BRIONES-GARCÍA, E.; CRUZ-DOMÍNGUEZ, M. P.; FLÓREZ-DURANTE, O. I.; JARA, L. J. Antiphospholipid antibodies disappearance in primary antiphospholipid syndrome: Thrombosis recurrence. **Autoimmunity Reviews**, v. 16, n. 4, p. 352-354, 2017.

MORELLI, V. M.; LIJFERING, W. M.; ROSENDAAL, F. R.; CANNEGIETER, S. C. Lipid levels and risk of recurrent venous thrombosis: results from the MEGA follow-up study. **Journal of Thrombosis and Haemostasis**, v. 15, n. 4, p. 695-701, 2017.

NIETO RODRÍGUEZ, J. A.; RAMÍREZ LUNA, J. C. Anticoagulant therapy duration. In favor of short-term courses. **Revista Clínica Española (English Edition)**, v. 217, n. 6, p. 365-369, 2017.

OKIN, P. M.; DEVEREUX, R. B.; FABSITZ, R. R.; LEE, E. T.; GALLOWAY, J. M.; HOWARD, B. V. Principal Component Analysis of the T Wave and Prediction of Cardiovascular Mortality in American Indians. **Circulation**, v. 105, n. 6, p. 714, 2002.

PABINGER, I.; SCHNEIDER, B. Thrombotic Risk in Hereditary Antithrombin III, Protein C, or Protein S Deficiency. **A Cooperative, Retrospective Study**, v. 16, n. 6, p. 742-748, 1996.

PALARETI, G.; COSMI, B.; LEGNANI, C.; ANTONUCCI, E.; DE MICHELI, V.; GHIRARDUZZI, A.; POLI, D.; TESTA, S.; TOSETTO, A.; PENGO, V.; PRANDONI, P. D-dimer to guide the duration of anticoagulation in patients with venous thromboembolism: a management study. **Blood**, v. 124, n. 2, p. 196-203, 2014.

PALARETI, G.; LEGNANI, C.; COSMI, B.; GUAZZALOCA, G.; PANCANI, C.; COCCHERI, S. Risk of venous thromboembolism recurrence: high negative predictive value of D-dimer performed after oral anticoagulation is stopped. **Thrombosis and haemostasis**, v. 87, n. 1, p. 7-12, 2002.

PALARETI, G.; LEGNANI, C.; COSMI, B.; VALDRÉ, L.; LUNGHI, B.; BERNARDI, F.; COCCHERI, S. Predictive Value of D-Dimer Test for Recurrent Venous Thromboembolism After Anticoagulation Withdrawal in Subjects With a Previous Idiopathic Event and in Carriers of Congenital Thrombophilia. **Circulation**, v. 108, n. 3, p. 313-318, 2003.

PARVEEN, R.; NABI, M.; MEMON, F.; ZAMAN, S.; ALI, M. A review and survey of artificial neural network in medical science. **J Adv Res Comput Appl**, v. 3, n. 1, p. 8-17, 2016.

PENGO, V.; RUFFATTI, A.; LEGNANI, C.; GRESELE, P.; BARCELLONA, D.; ERBA, N.; TESTA, S.; MARONGIU, F.; BISON, E.; DENAS, G.; BANZATO, A.; PADAYATTIL JOSE, S.; ILICETO, S. Clinical course of high-risk patients diagnosed with antiphospholipid syndrome. **Journal of Thrombosis and Haemostasis**, v. 8, n. 2, p. 237-242, 2010.

POWELL, M. J. D. Restart procedures for the conjugate gradient method. **Mathematical programming**, v. 12, n. 1, p. 241-254, 1977.

PRANDONI, P.; NOVENTA, F.; GHIRARDUZZI, A.; PENGO, V.; BERNARDI, E.; PESAVENTO, R.; IOTTI, M.; TORMENE, D.; SIMIONI, P.; PAGNAN, A. The risk of recurrent venous thromboembolism after discontinuing anticoagulation in patients with acute proximal deep vein thrombosis or pulmonary embolism. A prospective cohort study in 1,626 patients. **Haematologica**, v. 92, n. 2, p. 199-205, 2007.

QATAWNEH, Z.; ALSHRAIDEH, M.; ALMASRI, N.; TAHAT, L.; AWIDI, A. Clinical decision support system for venous thromboembolism risk classification. **Applied Computing and Informatics**, 2017.

REZENDE, S. M.; LIJFERING, W. M.; ROSENDAAL, F. R.; CANNEGIETER, S. Hematological variables and venous thrombosis: red cell distribution width and blood monocytes are associated with an increased risk. **Haematologica**, 2013.

RIEDMILLER, M.; BRAUN, H. **RPROP-A fast adaptive learning algorithm**. Proc. of ISCIS VII), Universitat: Citeseer, 1992a.

RIEDMILLER, M.; BRAUN, H. **RPROP-A fast adaptive learning algorithm**. Proceedings of ISCIS VII: Citeseer, 1992b.

RIVA, N.; BELLESINI, M.; DI MINNO, M. N. D.; MUMOLI, N.; POMERO, F.; FRANCHINI, M.; FANTONI, C.; LUPOLI, R.; BRONDI, B.; BORRETTA, V. Poor predictive value of contemporary bleeding risk scores during long-term treatment of venous thromboembolism. **Thrombosis and haemostasis**, v. 112, n. 03, p. 511-521, 2014.

RODGER, M. A.; KAHN, S. R.; WELLS, P. S.; ANDERSON, D. A.; CHAGNON, I.; LE GAL, G.; SOLYMOSS, S.; CROWTHER, M.; PERRIER, A.; WHITE, R.; VICKARS, L.; RAMSAY, T.; BETANCOURT, M. T.; KOVACS, M. J. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. **CMAJ : Canadian Medical Association Journal**, v. 179, n. 5, p. 417-426, 2008.

RODGER, M. A.; LE GAL, G.; ANDERSON, D. R.; SCHMIDT, J.; PERNOD, G.; KAHN, S. R.; RIGHINI, M.; MISMETTI, P.; KEARON, C.; MEYER, G.; ELIAS, A.; RAMSAY, T.; ORTEL, T. L.; HUISMAN, M. V.; KOVACS, MICHAEL J. Validating the HERDOO2 rule to guide treatment duration for women with unprovoked venous thrombosis: multinational prospective cohort management study. **BMJ**, v. 356, 2017.

ROSENDAAL, F. R. Causes of venous thrombosis. **Thrombosis Journal**, v. 14, n. 1, p. 24, 2016.

ROSENDAAL, F. R.; REITSMA, P. H. Genetics of venous thrombosis. **Journal of Thrombosis and Haemostasis**, v. 7, p. 301-304, 2009.

RUPA-MATYSEK, J.; GIL, L.; WOJTASIŃSKA, E.; CIEPŁUCH, K.; LEWANDOWSKA, M.; KOMARNICKI, M. The relationship between mean platelet volume and thrombosis recurrence in patients diagnosed with antiphospholipid syndrome. **Rheumatology International**, v. 34, n. 11, p. 1599-1605, 2014.

SAHA, P.; MANDAL, R. Detection of Dengue Disease Using Artificial Neural Networks. 2017.

SARAIVA, S. D. S.; CUSTÓDIO, I. F.; MAZETTO, B. D. M.; COLLELA, M. P.; DE PAULA, E. V.; APPENZELLER, S.; ANNICHINO-BIZZACHI, J.; ORSI, F. A. Recurrent thrombosis in antiphospholipid syndrome may be associated with cardiovascular risk factors and inflammatory response. **Thrombosis Research**, v. 136, n. 6, p. 1174-1178, 2015.

SASAKI, K.; KANTARJIAN, H. M.; JABBOUR, E. J.; O'BRIEN, S.; RAVANDI, F.; KONOPLEVA, M.; BORTHAKUR, G.; WIERDA, W. G.; DAVER, N.; TAKAHASHI, K.; JAIN, P.; SKINNER, J.; RIOS, M. B.; PIERCE, S.; GARCIA-MANERO, G.; CORTES, J. E. Clinical Application of Artificial Intelligence in Patients with Chronic Myeloid Leukemia in Chronic Phase. **Blood**, v. 128, n. 22, p. 940-940, 2016.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural Networks**, v. 61, p. 85-117, 2015.

SCHULMAN, S.; GRANQVIST, S.; HOLMSTRÖM, M.; CARLSSON, A.; LINDMARKER, P.; NICOL, P.; EKLUND, S.-G.; NORDLANDER, S.; LÄRFARS, G.; LEIJD, B. The duration of oral anticoagulant therapy after a second episode of venous thromboembolism. **New England Journal of Medicine**, v. 336, n. 6, p. 393-398, 1997.

SHLIPAK, M. G.; FRIED, L. F.; CRUMP, C.; BLEYER, A. J.; MANOLIO, T. A.; TRACY, R. P.; FURBERG, C. D.; PSATY, B. M. Elevations of Inflammatory and Procoagulant Biomarkers in Elderly Persons With Renal Insufficiency. **Circulation**, v. 107, n. 1, p. 87-92, 2003.

SILVERSTEIN, M. D.; HEIT, J. A.; MOHR, D. N.; PETTERSON, T. M.; O'FALLON, W.; MELTON, L.; III. Trends in the incidence of deep vein thrombosis and pulmonary embolism: A 25-year population-based study. **Archives of Internal Medicine**, v. 158, n. 6, p. 585-593, 1998.

SIRAGUSA, S.; MALATO, A.; ANASTASIO, R.; CIGNA, V.; MILIO, G.; AMATO, C.; BELLISI, M.; ATTANZIO, M. T.; CORMACI, O.; PELLEGRINO, M.; DOLCE, A.; CASUCCIO, A.; BAJARDI, G.; MARIANI, G. Residual vein thrombosis to establish duration of anticoagulation after a first episode of deep vein thrombosis: the Duration of Anticoagulation based on Compression UltraSonography (DACUS) study. **Blood**, v. 112, n. 3, p. 511-515, 2008.

SÜT, N.; ÇELIK, Y. Prediction of mortality in stroke patients using multilayer perceptron neural networks. **Turkish Journal of Medical Sciences**, v. 42, n. 5, p. 886-893, 2012.

THORPE, M. G.; MILTE, C. M.; CRAWFORD, D.; MCNAUGHTON, S. A. A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. **International Journal of Behavioral Nutrition and Physical Activity**, v. 13, n. 1, p. 30, 2016.

TOSETTO, A.; IORIO, A.; MARCUCCI, M.; BAGLIN, T.; CUSHMAN, M.; EICHINGER, S.; PALARETI, G.; POLI, D.; TAIT, R. C.; DOUKETIS, J. Predicting disease recurrence in patients with previous unprovoked venous thromboembolism: a proposed prediction score (DASH). **Journal of Thrombosis and Haemostasis**, v. 10, n. 6, p. 1019-1025, 2012.

TOSETTO, A.; TESTA, S.; MARTINELLI, I.; POLI, D.; COSMI, B.; LODIGIANI, C.; AGENO, W.; DE STEFANO, V.; FALANGA, A.; NICHELE, I.; PAOLETTI, O.; BUCCIARELLI, P.; ANTONUCCI, E.; LEGNANI, C.; BANFI, E.; DENTALI, F.; BARTOLOMEI, F.; BARCELLA, L.; PALARETI, G. External validation of the DASH prediction rule: a retrospective cohort study. **Journal of Thrombosis and Haemostasis**, v. 15, n. 10, p. 1963-1970, 2017.

TRITSCHLER, T.; MÉAN, M.; LIMACHER, A.; RODONDI, N.; AUJESKY, D. Predicting recurrence after unprovoked venous thromboembolism: prospective validation of the updated Vienna Prediction Model. **Blood**, v. 126, n. 16, p. 1949-1951, 2015.

TSAI, A. W.; CUSHMAN, M.; ROSAMOND, W. D.; HECKBERT, S. R.; POLAK, J. F.; FOLSOM, A. R. Cardiovascular risk factors and venous thromboembolism incidence: The

longitudinal investigation of thromboembolism etiology. **Archives of Internal Medicine**, v. 162, n. 10, p. 1182-1189, 2002.

VAN DER HULLE, T.; TAN, M.; DEN EXTER, P. L.; VAN ROOSMALEN, M. J. G.; VAN DER MEER, F. J. M.; EIKENBOOM, J.; HUISMAN, M. V.; KLOK, F. A. Recurrence risk after anticoagulant treatment of limited duration for late, second venous thromboembolism. **Haematologica**, v. 100, n. 2, p. 188-193, 2015.

VAVOUGIOS, G. D.; NATSIOS, G.; PASTAKA, C.; ZAROGIANNIS, S. G.; GOURGOULIANIS, K. I. Phenotypes of comorbidity in OSAS patients: combining categorical principal component analysis with cluster analysis. **Journal of Sleep Research**, v. 25, n. 1, p. 31-38, 2016.

VAYÁ, A.; SUESCUN, M. Hemorheological parameters as independent predictors of venous thromboembolism. **Clinical hemorheology and microcirculation**, v. 53, n. 1-2, p. 131-141, 2013.

VEINGUIDE.COM. **Vein Valves and Blood Clots, Venous Blood Clots, Venous Valves and Blood Clots**. Disponível em: <http://www.veinguide.com/blog/257/vein-valves-and-blood-clots-venous-blood-clots-venous-valves-and-blood-clots-on-www.veinguide.com-and-www.santamonicaveincenter.com.aspx>. Acesso em: 01/04/2018

VERHOVSEK, M.; DOUKETIS, J. D.; YI, Q.; ET AL. Systematic review: D-dimer to predict recurrent disease after stopping anticoagulant therapy for unprovoked venous thromboembolism. **Annals of Internal Medicine**, v. 149, n. 7, p. 481-490, 2008.

VUČKOVIĆ, B. A.; CANNEGIETER, S. C.; VAN HYLCKAMA VLIEG, A.; ROSENDAAL, F. R.; LIJFERING, W. M. Recurrent venous thrombosis related to overweight and obesity: results from the MEGA follow-up study. **Journal of Thrombosis and Haemostasis**, v. 15, n. 7, p. 1430-1435, 2017.

WELLS, R. E.; MERRILL, E. W. Influence Of Flow Properties Of Blood Upon Viscosity-Hematocrit Relationships. **Journal of Clinical Investigation**, v. 41, n. 8, p. 1591-1598, 1962.

WHITE, R. H.; ZHOU, H.; ROMANO, P. S. Length of hospital stay for treatment of deep venous thrombosis and the incidence of recurrent thromboembolism. **Archives of Internal Medicine**, v. 158, n. 9, p. 1005-1010, 1998.

XU, Z.; LIOI, J.; MU, J.; KAMOCKA, M. M.; LIU, X.; CHEN, D. Z.; ROSEN, E. D.; ALBER, M. A Multiscale Model of Venous Thrombus Formation with Surface-Mediated Control of Blood Coagulation Cascade. **Biophysical Journal**, v. 98, n. 9, p. 1723-1732, 2010.

YU, F. T. H.; ARMSTRONG, J. K.; TRIPETTE, J.; MEISELMAN, H. J.; CLOUTIER, G. A local increase in red blood cell aggregation can trigger deep vein thrombosis: evidence based on quantitative cellular ultrasound imaging. **Journal of Thrombosis and Haemostasis**, v. 9, n. 3, p. 481-488, 2011.

ZAGO, M. A.; FALCAO, R. P.; PASQUINI, R. **Hematologia: fundamentos e prática**. Atheneu, 2004.

ZHU, D. Mathematical modeling of blood coagulation cascade: kinetics of intrinsic and extrinsic pathways in normal and deficient conditions. **Blood Coagulation & Fibrinolysis**, v. 18, n. 7, p. 637-646, 2007.

ZHU, T.; MARTINEZ, I.; EMMERICH, J. Venous Thromboembolism. **Risk Factors for Recurrence**, v. 29, n. 3, p. 298-310, 2009.