Universidade do Minho
Escola de Engenharia

Beatriz Teixeira de Magalhães

Development of a Scoring System to Assess Potential Biomarkers for Atrial Fibrillation

Tese de Mestrado
Mestrado em Bioinformática

Trabalho efetuado sob a orientação do
Professor Doutor Rui Vitorino
Professor Doutor Miguel Rocha

Julho de 2018

Universidade do Minho
Escola de Engenharia

Beatriz Teixeira de Magalhães

Development of a Scoring System to Assess Potential Biomarkers for Atrial Fibrillation

Tese de Mestrado
Mestrado em Bioinformática

Trabalho efetuado sob a orientação do
Professor Doutor Rui Vitorino
Professor Doutor Miguel Rocha

Julho de 2018

DECLARAÇÃO

Nome: Beatriz Teixeira de Magalhães
Endereço eletrónico: pg32936@alunos.uminho.pt Telefone: 914801960
Número do Bilhete de Identidade: 14752848
Título dissertação/tese: Development of a Scoring System to Assess Potential Biomarkers for Atrial Fibrillation

Orientadores:
Professor Doutor Rui Vitorino, Professor Doutor Miguel Rocha

Ano de conclusão: 2018
Designação do Mestrado ou do Ramo de Conhecimento do Doutoramento:
Mestrado em Bioinformática – Tecnologias de Informação

Universidade do Minho, 17/07/2018

Assinatura: Beatriz Teixeira de Magalhães

Acknowledgements/ Agradecimentos

Ao Prof. Rui, por ter acreditado em mim e por todas as oportunidades que me tem facultado. Obrigada ainda pela excelente orientação, apoio e, sobretudo, pela confiança. Espero não ter desapontado!

Ao Prof. Miguel, por ter sido um excelente professor e o principal responsável por todos os conhecimentos que adquiri na área da programação. Obrigada por toda a disponibilidade demonstrada quer durante a realização desta dissertação, quer durante todo o mestrado.

Ao Fábio, porque todo este trabalho começou com ele e não teria sido realizado de outra forma. Agradeço ainda toda a disponibilidade, ajuda e preocupação que sempre demonstrou.

À Marta, ao João e ao Daniel porque estes dois anos não teriam sido os mesmos sem vocês. Tornaram todo este percurso mais fácil, pela companhia, pela entreajuda e pelo apoio. Sem dúvida alguma que foram parte indispensável e não poderia ter pedido melhores pessoas para viver esta etapa comigo.

À Inês, à Lisa, à Ritinha, à Caty e à Rita por estarem sempre lá e por serem as melhores amigas que alguém poderia ter. Por todos os risos, brincadeiras e conversas e por me fazerem esquecer o trabalho que ainda falta fazer e descomprimir pelo menos por alguns momentos. São amigas para a vida e tenho a certeza de que vou sempre poder contar com elas e elas comigo. Ao João Luís que é sem dúvida a pessoa mais criativa e talentosa que conheço e cujas exposições me alegram sempre, para além, claro, do seu sentido de humor sempre presente.

Ao André por ser um dos meus pilares de apoio, por estar sempre ao meu lado, por acreditar em mim, por me incentivar e especialmente por ser quem é. Obrigada por me aturares e por me compreenderes.

À Ana, porque eu não seria a mesma sem ela, porque me ensinou a ser quem sou, porque me ensina a ser melhor e sempre a manter os pés assentes na terra. Contigo aprendo constantemente e, apesar de nunca to ter dito, és, sempre foste e sempre serás o meu ídolo e a minha maior inspiração e aspiração. À Clumsy por simplesmente existir, por me fazer esquecer tudo só de olhar para ela e pela companhia de todos os dias.

Aos meus pais, as pessoas mais importantes, porque sem eles não estaria aqui, porque foram os primeiros a acreditar em mim, porque sempre me apoiaram, porque possibilitaram todo o meu percurso académico desde a escola primária até aqui. Obrigada por serem os melhores pais do mundo, obrigada por nunca duvidarem de mim, obrigada por me fazerem lutar pelos meus sonhos.

Por fim, obrigada ao resto da minha família por todo o apoio.

Development of a Scoring System to Assess Potential Biomarkers for Atrial Fibrillation

Abstract

Atrial fibrillation affects millions of individuals worldwide, posing a major threat to public health due to the variety of comorbidities that constitute by-products of the disease. In light of this epidemic, new means of diagnosis, prognosis and therapy are pressing. Biomarkers, particularly protein markers, are important tools in this process but lack validation, which is essential before clinical translation. Several appraisal benchmarks have been developed to determine the relative potential of biomarkers, but these present multiple limitations.

We developed a bioinformatic-oriented scoring function aimed at weighing the importance of proteins and mitigating the limitations of the currently known scores. After taking an extensive literature search and mining a massive volume of reports, data was organized into several subsets, according to the sample major characteristic and atrial fibrillation type. A mathematical scoring function was proposed, based on the consensus of studies supporting the protein-disease association (incoherence), median of the reported fold-changes and importance of each study according to the number of diseased individuals, and applied to each subset in the form of an algorithm implemented in Python 3.5.

The developed ranking method performed well regarding both the degree of alteration and the inconsistency parameters. Our results portray a set of proteins with the highest biomarker potential (highest scores) for atrial fibrillation. We also selected the top five potential biomarkers for atrial fibrillation in general and for each type of disease. The main biological functions in which they are involved were retrieved for comparison with the state of the art. Alterations in the expression levels of proteins involved in either of these functions seem to agree with AF's pathophysiology and clinical presentation, showing the effectiveness of the developed algorithm.

Overall, the developed pipeline seems to improve the processes of biomarker ranking and selection for a target disease, allowing a leap towards clinical translation.

Desenvolvimento de um Sistema de Classificação para Aferir Potenciais Biomarcadores para a Fibrilhação Auricular

Resumo

A fibrilhação auricular afeta milhões de indivíduos em todo o mundo, representando uma grande ameaça à saúde pública devido à grande variedade de comorbidades que constituem subprodutos da mesma. Face a esta epidemia, novos métodos de diagnóstico, prognóstico e terapêutica são prementes. Os biomarcadores, em particular marcadores proteicos, tornam-se importantes ferramentas neste processo, mas carecem de validação, passo essencial antes da tradução clínica. Vários meios de avaliação foram desenvolvidos para determinar o seu potencial relativo, mas estes apresentam inúmeras limitações.

Neste trabalho desenvolvemos uma função de classificação orientada para a bioinformática, destinada a calcular a importância de proteínas e a mitigar as limitações dos métodos já conhecidos. Após uma extensa pesquisa de literatura e análise de um volume enorme de artigos, os dados foram organizados em vários subconjuntos, de acordo com a principal característica da amostra e tipo de fibrilhação auricular. Uma função matemática de classificação foi proposta, baseada no consenso de estudos que suportam a associação proteína-doença (incoerência), mediana dos *fold-changes* e importância de cada estudo de acordo com o número de indivíduos afetados, e aplicada a cada subconjunto por meio de um algoritmo implementado em *Python* 3.5.

O método de classificação desenvolvido teve uma boa performance relativamente a ambos os parâmetros, nomeadamente o grau de alteração e a coerência. Os resultados retratam um conjunto de proteínas com o potencial de biomarcador mais elevado (classificações mais elevadas) para a fibrilhação auricular. Também selecionamos as cinco proteínas com o maior potencial de biomarcador para a fibrilhação auricular geral e para cada tipo da doença. Procedeu-se um rastreio das principais funções biológicas nas quais as proteínas estão envolvidas para comparação com o estado da arte. Alterações nos níveis de expressão de proteínas envolvidas em qualquer uma destas funções parecem estar de acordo com a patofisiologia e apresentação clínica desta arritmia, o que demonstra a eficácia do algoritmo desenvolvido.

De forma geral, todo o processo aqui delineado parece melhorar os processos de classificação e seleção de biomarcadores para uma doença alvo, permitindo progressos na direção da tradução clínica.

Index

List of Figures

List of Tables

List of Abbreviations and Acronyms

1D SDS-PAGE - One(1)-Dimensional SDS-Polyacrilamide Gel Electrophoresis

2DE - Two(2)-Dimensional Gel Electrophoresis

ACTC1 - Actin, alpha cardiac muscle 1

AF - Atrial Fibrillation

ANP - Atrial Natriuretic peptide

APLN12 - Apelin-12

ATF6 - Activating Transcription Factor 6

ATR - Atrial Tachycardia Remodelling

B-TGF - Beta-thromboglobulin

BNP - Brain Natriuretic Peptide

CHI3L1 - Chitinase 3-like protein 1

CREB/ATF - cAMP response element binding protein, activating transcription factors

CRP - C-reactive protein

CST-C - Cystatin-C

Ca^{2+} - Calcium Ion

DAD - Delayed Afterdepolarization

DCSM - Document Conversion and Structuring Module

DD - D-dimer

DES - Desmin

DGS - Direção Geral de Saúde

DMQH - 5-demethoxyubiquinone hydroxylase, mithocondrial

DRM - Document Retrieval Module

E2F1 - E2F Transcription Factor

EAD - Early Afterdepolarization

ECG - Electrocardiogram

ELISA - Enzyme-Linked Immunosorbent Assay

ESC - European Society of Cardiology

GDA - Gene-Disease Association

GDF-15 - Growth/differentiation factor 15

GOBIOM - Global Online Biomarker Database

HATCH - Heart Failure, Age, Previous Transient Ischaemic Attack or Stroke, Chronic Obstructive Pulmonary Disease, Hypertension

HBA1 - Hemoglobin subunit alpha

HGF - Hepatocyte growth factor

HMDB - The Human Metabolome Database

HSF1 - Heat Shock Transcription Factor 1

HSP - Heat Shock Protein

HSPA5 - Endoplasmic reticulum chaperone BiP

IDF - Inverse Document Frequency

IE - Information Extraction

IL-10 - Interleukin-10

IL-18 - Interleukin-18

IR - Information Retrieval

K^+ - Potassium Ion

LAA - Left Atrial Appendage

LFC - Log Fold Change

LS_d - Differential Link Score

LS_i - Interaction Link Score

M-CK - Creatine Kinase M type

MMP-9 - Matrix metalloproteinase 9

MRproANP - Mid-regional pro-Atrial Natriuretic peptide

MS - Mass Spectrometry

NCX - Ca^{2+}/Na^+ Exchanger

NF - Normalized Frequency Association

NLPM - Natural Language Processing Module

NTANP - N-terminal Atrial Natriuretic peptide

NTproBNP - N-terminal pro-Brain Natriuretic Peptide

Na^+ - Sodium Ion

PBP - Platelet basic protein’s cleavage

PF-4 - Platelet factor 4

PTX3 - Pentraxin-related protein PTX3

RAA - Right Atrial Appendage

RyRs - Ryanodine Receptors

SAA1 - Serum amyloid A-1 protein

SERCA - Sarcoplasmic Reticulum Ca^{2+} ATPase

SPR - Sarcoplasmic Reticulum

SR - Sinus Rhythm

SRF - Serum Response Factor

TF - Tissue Factor

TGF-B-1 - Transforming Growth factor beta 1

TH3 - Triiodothyronine

TMM - Text-mining Module

TNFRSF11A - Tumor necrosis factor receptor superfamily member 11A

TNFSF11 - Tumor necrosis factor ligand superfamily member 11

TnIc - Troponin I, cardiac muscle

U-II - Urotensin-2

VCAM-1 - Vascular cell adhesion protein 1

VDA - Variant-Disease Association

VEGF-A - Vascular endothelial growth factor A

VEGFR-1 - Vascular endothelial growth factor receptor 1

t-PA - Tissue-type plasminogen activator

vWF - von Willebrand factor

1. Introduction

1.1 Motivation

Atrial fibrillation (AF) is a cardiac rhythm disturbance associated with high cardiovascular morbidity and mortality rates [1]. The prevalence and incidence of AF increase with advancing age [2], making it a particular problem among the elderly. Thus, preventive strategies to identify those who are at risk of developing the disease are of utmost importance. The association of AF with adverse outcomes, including stroke, prompts the need to prevent it. As such, it is necessary to know the conditions that predispose to episodes of arrhythmia and its prognosis [3]. There is also a recognized need to improve AF's detection rates prior to the development of first complications and to provide adequate anticoagulation therapy to all eligible patients [4]. Single time-point screening with electrocardiogram (ECG) or pulse palpitation measurement have shown multiple benefits in older patients. However, these are not very effective in screening paroxysmal AF, given that subjects may not be experiencing an AF episode at that time. Additionally, pulse palpitation has a high-degree of sensitivity for AF (87%-97%) but there is an element of subjectivity that may lead to false positives (specificity range 70%-81%)[5]. To develop effective screening strategies for AF, appropriate target populations and timing must be considered to improve the chances of identifying paroxysmal patients in the most convenient and cost-effective manner for primary care [6].

Multiple biological markers that have the ability to predict the future development of AF have been identified. Such markers allowed a better knowledge of AF's pathophysiology, giving light to several processes that either initiate or perpetuate the disease. Importantly, they may give prognostic information [7], allowing the anticipation of corrective therapies. This thesis will focus on protein biomarkers, which can be used for prognostic or diagnostic purposes or even as pharmacological targets. Introducing the use of biomarkers into the clinical practice would facilitate these processes, which could be achieved with a procedure as simple as urine or blood sample measurement of the specific markers' levels. The challenge nonetheless, is to determine which proteins are differentially expressed in a disease, in this case AF, and if there is consensus between different studies. When validated, that is when a protein is associated with a certain disease and there is a high-degree of consensus between the majority of publications and independent experimental evidence, a protein could be considered a biomarker. To address this issue, objective scoring algorithms can be useful to unbiasedly sort and pinpoint surrogate markers retrieved from the literature. Hence, a scoring method was developed and applied to datasets created from text-mining, with the ultimate aim of highlighting differentially expressed proteins in patients with AF and determining a subset of those proteins that could represent potential biomarkers for this condition.

1.2 Objectives

The main aim of this work was to design biomarker panels for atrial fibrillation diagnosis and prognosis, based on a scoring function defined over protein quantification data. In detail, the scientific and technological objectives were to:

1) Retrieve and manually curate single protein-centred and proteomics-based literature to create a dataset/database of proteins that may or may not be altered in AF;
2) Create/formulate a mathematical scoring function which weighs the importance of proteins as potential biomarkers for AF;
3) Design and implement an algorithm applying the created scoring function to the protein dataset collected;
4) Define panels of potential biomarkers for AF and its several types as proof of concept;
5) Determine the uncertainty of the results obtained by the developed scoring approach using a bootstrap-based system.

1.3 Structure of the Document

The present document is organized as follows:

Chapter 2 – State of The Art

Introduction to atrial fibrillation, namely characterization, classification, epidemiology, risk factors, means of diagnosis, management and prognosis and pathophysiology. Concept of the term biomarker, importance of their use and advances regarding its study or use in atrial fibrillation. Discussion of the importance of text-mining and bioinformatics in the field of biomarkers, comparison between manual and automated text-mining and mention and appraisal of existing ranking methods for biomarkers.

Chapter 3 – Methods

Description of the extraction and retrieval of relevant literature processes, the developed scoring function and the validation steps as part of the followed pipeline.

Chapter 4 – Results

Presentation of the main results obtained.

Chapter 5 – Discussion

Discussion of the performance of the created approach and of the results.

Chapter 6 – Conclusion and Future Work

Main conclusion of the developed work and results and description of possible future endeavours and applications of the conceived algorithm.

2. State of the Art

2.1 Characterization, Classification and Epidemiology of Atrial Fibrillation

Atrial fibrillation (AF) is characterized by the rapid and irregular activation of the atria, 400-600 pulses per minute. In normal conditions, the heart rate is adjusted according to the body's metabolic needs through physiological control of the sinoatrial node, which maintains a rhythm of around 60 beats per minute at rest and up to 180-200 beats per minute during exercise. Instead, atrial cells of a patient with AF fire at a much higher rate, which would lead to ineffective cardiac contraction and rapid death if conducted to the ventricles. The atrioventricular node acts as a filter or obstacle, with limited impulse-carrying capacity, through which the atrial impulses must pass before activating the ventricles. Hence, the ventricular rate during AF is no longer controlled by the sinoatrial node, but by the interaction between the atrial rate and the atrioventricular node. In the absence of any drug therapy, the ventricular rate in AF patients is of about 150 pulses per minute [8].

According to the 2016 European Guidelines for the management of AF, the arrhythmia can be classified into five types: first-diagnosed, if it has not been diagnosed before, regardless of the duration or the presence and severity of AF-related symptoms; paroxysmal, when it lasts less than 7 days and spontaneously converts to sinus rhythm (SR) or is cardioverted; persistent, if it occurs for a period of at least seven days; long-standing persistent, if it lasts for more than one year; and permanent, when no further attempts to return to SR are made [4]. Individuals with long-standing persistent AF tend to be older and have more comorbidities [9]. The different types of AF are summarized in **Table 1**.

Guidelines state that patients showing arrhythmia lasting for at least seven days have the persistent form, but individuals with AF of longer duration are more likely to have sustained greater extent of atrial remodelling and respond poorer to long-term treatment compared to the previous. The American College of Cardiology and the American Heart Association also consider an additional type of AF known as lone AF. Individuals with this type of arrhythmia are younger, present no cardiopulmonary disease nor hypertension and face a lower risk of thromboembolism in earlier stages. Notwithstanding, as the disease progresses with age, the risk of thromboembolism and mortality increases and, subsequently, patients may respond differently to the same treatment [10]. Whether the term lone AF should actually be used is questionable. Some guidelines do provide a definition of this type of AF but do not provide direction about how much or what kind of tests are warranted to exclude heart disease. Moreover, there is variability in the definitions encountered in the literature, and so the term should be avoided [11].

Each year, approximately 5% of paroxysmal AF patients progress to the persistent form [12]. Progression from persistent to permanent AF has an even bigger rate, with 35% to 40% of the patients progressing in less than a year [13]. In younger patients with lone AF the rate of progression is lower, only 1% to 3% per year [14]. Still, AF's progression presents a wide variability among individuals. For instance, in new onsets of the disease the presenting form is persistent AF [12]. In a study conducted along 30 years, the probability of progression from paroxysmal or persistent to permanent AF was lower than expected. AF can induce electrophysiological changes that tend to perpetuate the arrhythmia, but the results from the same study indicate that in individuals without heart disease, these proarrhythmic effects are insufficient for progression in the absence of comorbidities [14].

Table 1 – Characterization of Atrial Fibrillation.

Type of AF	Definition	Reference
First-diagnosed	AF that has not been diagnosed before.	Kirchhof P. *et al*. (2016)
Paroxysmal	AF lasting < 7 days and spontaneously or through cardioversion reverts to SR. Caused by focal drivers, especially in the cardiomyocytes sleeves around the pulmonary veins.	Kirchhof P. *et al*. (2016); Burstein B. *et al*. (2008)
Persistent	AF lasting ≥ 7 days, including episodes terminated by cardioversion after at least 7 days. Caused by functional re-entry substrates.	Kirchhof P. *et al*. (2016); Nattel S. *et al*. (2011)
Long-standing persistent	AF lasting for ≥ 1 year.	Kirchhof P. *et al*. (2016)
Permanent	AF that is accepted by the patient and the physician. No further attempts to return to SR are made. It occurs when the substrate becomes fixed and irreversible due to structural remodelling.	Kirchhof P. *et al*. (2016); Allessie M.A. *et al*. (2001)

AF = Atrial Fibrillation
SR = Sinus Rhythm

AF is the most common type of arrhythmia encountered in clinical practice with a prevalence set to increase due to ageing trends in the global population [15]. In fact, the current demographic transition to an inverted pyramid age may partly explain the also rising incidence of AF [16]. Nonetheless, comorbidities, cardiovascular risk factors and lifestyle changes may also lie on the root of this growth [17,18]. Both the incidence and prevalence of AF double with each passing decade for people older than 50 and reach 10% in octogenarians [3]. In Portugal, according to the Direção Geral de Saúde (DGS) the prevalence of AF is not clearly defined [19]. In 2003, a study conducted with patients from Portugal's health centres found a prevalence of 0.53% in men and 0.54% in women. Such prevalence raised progressively with age: 0.02% in individuals with 35-44 years old, 0.13% in individuals with 45-54 years old, 0.63% between 55-64 years old, 1.83% between 65-74 years old, reaching 2.87% in individuals with 75 or more years old [20]. However, more recently a prevalence of 2.5% in both men and women in Portugal was described [21], clearly indicating that levels have been rising in the last decade.

The burden of AF in public health was measured as disability-adjusted life-years by S. Chugh and collaborators (2014) [16], who found an increase of 18.8% in men and 18.9% in women from 1990 to 2010. Such finding was also sided by a rise in both men and women mortality rates in the same period. Truly, AF tends to double the risk of mortality from cardiovascular and other causes [22].

AF poses a major risk factor for stroke, with considerable weight regardless of the individual's age. One may hypothesize that coexisting conditions (e.g.: coronary artery disease) are the cause of the excess of strokes rather than AF itself. Nevertheless, observations from the Framingham study show that the occurrence of stroke in people with coronary artery disease was higher in individuals with concomitant AF and the recurrence of stroke happened earlier in AF patients [23].

In the general population, the risk of stroke is inferior to 1.7-2.1% according to Portuguese data ([24–30] cited by [20]). In the presence of AF, the risk increases to more than 5% per year, even without further risk factors [31,32]. This augmented risk is explained by the loss of atrial contraction which leads to stasis of the blood in the atria. Such stasis promotes clot formation and thromboemboli, which tend to propagate to other organs, particularly the brain [8]. Recent data from Portugal also shows that 25% of patients with ischaemic cerebral vascular accidents and subjected to intravenous fibrinolysis in the first three hours of onset of signs exhibit AF. Furthermore, it was observed that 22% of the ischaemic patients presented cardioembolism, with AF being responsible for 80% of these events [19]. Moreover, sustained AF, along with an uncontrolled ventricular rate, can cause severe congestive heart failure, although reversible by proper rate or rhythm control [33].

AF is also associated with significantly impaired quality of life (e.g.: frequent hospitalization) [16]. Such impact has socio-economic repercussions worldwide owing to hospital admissions, chronic disease management and disabilities [18]. Therefore, an understanding of the underlying mechanisms of this arrhythmia is crucial to the identification of new therapeutic drugs and to prevent or subside the progression of AF [15]. Additionally, the present pathophysiological insights of AF suggest that early diagnosis and comprehensive therapy could help in preventing progression, reducing AF-related complications [34].

2.2 Risk Factors

Many factors and conditions are thought to predispose the development of AF. AF's episodes are initiated by a trigger acting on a vulnerable substrate, at least partially determined by genetic factors [35,36]. In the absence of risk factors, several mutations and gene variants allow AF's initiation. Deshmukh *et al.* (2015) [37] contributed to the discovery of new genetic variants associated with AF. In this study, AF's susceptibility was related with decreased expression of the targets of cAMP response element binding protein, activating transcription factors (CREB/ATF) family, heat shock transcription factor 1 (HSF1), activating transcription factor 6 (ATF6), serum response factor (SRF), and E2F transcription factor 1 (E2F1) and persistent AF was associated with decreased expression in genes and gene sets related to ion channel function, consistent with reported functional changes.

Many genetic components are yet to be uncovered, but large population studies show that the risk of AF is doubled if there is parental history of arrhythmia [38]. Additional risk factors that develop overtime, combined with physiological aging or cardiac remodelling, make way for an appropriate trigger to initiate AF [9]. Genetic variants that increase the liability of AF's risk factors may, therefore, also raise the risk of AF [39].

Cardiovascular causes associated with AF include valvular heart disease, acute myocardial infarction, myocarditis, hypertrophic cardiomyopathy, congenital heart disease, pericarditis, hypertensive cardiovascular disease and heart failure [3]. Accordingly, in the Framingham study, cardiovascular diseases such as heart failure, myocardial infarction and valvular heart disease were found to be the most common pathological precursors of AF, accounting for 20% and 31% of AF's incidence in men and women, respectively [2]. As a matter of fact, 5% to 10% of patients with myocardial infarction [40,41] and up to 40% of patients that underwent cardiothoracic surgery [42] develop AF. Left-ventricular hypertrophy and hypertension were also established as significant AF predictors [3].

Non-cardiac risk factors linked to AF include thyrotoxicosis, alcohol abuse, severe infections, pulmonary pathology, smoking habits and diabetes [3].

Hypertension and ischaemic heart disease are the most common clinical settings on permanent AF. Additionally, if this subset of patients has congestive heart disease, the probability of developing the arrhythmia is even higher [43]. Notwithstanding, there is no obvious clinical cause for half of the patients with paroxysmal AF and for less than 20% of the patients with permanent or persistent forms [44]. The absence of identifiable predisposing factors is quite troublesome in that it makes targeting preventive therapy difficult [45], which turns the investigation of reliable diagnostic and prognostic tools imperative.

2.3 Diagnosis, Management and Prognosis

More than thirty million people worldwide were estimated to suffer from AF in 2010 [16]. An early and accurate diagnosis is crucial to provide anticoagulation therapy, which may prevent an initial ischaemic stroke event [4]. Unfortunately, in many cases, diagnosis occurs after a stroke event has taken place [46]. The biggest diagnostic challenge concerns paroxysmal or asymptomatic AF, especially according to the ASSERT and Copenhagen Holter studies, which indicate that even episodes of silent AF are associated with an increased risk of stroke [47,48].

Many benefits have been reported regarding the use of an ECG or pulse palpitation for single time-point screening of AF in older patients [49]. In the elderly study (SAFE, 2007) [50], active screening (opportunistic or systematic) among patients aged 65 or more was found to be more effective in detecting AF than routine care. In 2015, the European Primary Care Cardiovascular Society recommended opportunistic screening through the same approaches or by using modified sphygmomanometers or single-lead ECG devices if subject to independent validation with a 12-lead ECG [51].

In 2016, the previous European Society of Cardiology (ESC) recommendations, that is opportunistic screening through pulse palpitation followed by a confirmatory ECG, to all patients with 65 or more years old [52,53], were reiterated and systematic ECG screening was recommended in patients with more than 75 years or with high stroke risk. Extended screening in patients after a transient ischaemic attack or ischaemic stroke was also proposed and should include a short-term ECG followed by at least 72 hours of ECG monitoring. Using non-invasive ECG monitors or implanted loop recorders that allow long-term monitoring to detect silent AF episodes may be considered for stroke patients. Moreover, ESC guidelines recommended interrogating pacemakers and implantable cardioverter defibrillators on a regular basis to detect atrial high-rate episodes, which should trigger further investigation by ECG to document AF [4].

Even though pulse palpitation is a sensitive method for AF's screening, it lacks specificity. Accordingly, suspected AF should always be confirmed with 12-lead ECG. A normal ECG, however, does not rule out the diagnosis of AF because the patient might not be experiencing an episode at that time [11].

Non-pharmacological approaches to treat AF aim to modify the anatomical substrate causative of the arrhythmia or eliminate the trigger that initiates the same [54]. In the early 90's, Cox developed the Maze procedure, that proved to be very effective in treating AF ([55] cited by [54]). This technique is a relatively complex one. It consisted of an extensive dissection of right and left atrium, creating a sort of a maze through which the electrical activation was forced. This prevented the formation and perpetuation of the multiple wavelets that maintain AF [56]. Surgical dissection was eventually substituted by lesions provoked by different sources of energy like radiofrequency [57] or cryothermy [58]. All novel techniques have in common the fact that the posterior part of the left atrium and the pulmonary veins are involved in the ablation [54]. Thereby, nowadays there are essentially two ways of achieving rate control in patients with AF, pharmacologic treatment and ablation.

Antiarrhythmic drugs are seen as first-line treatment for rhythm control. They pose, however, proarrhythmic toxicity and have little efficacy in accomplishing and maintaining SR. Amiodarone seems to be the most effective pharmaceutical and yet, in trials in which it most successfully maintained SR, AF recurred in 35% of the cases [59]. Furthermore, antiarrhythmic drugs are not specific for atrial electrical activity and can have profound effects on ventricles [8], potentially leading to proarrhythmia and increased mortality [60]. Consequently, safer and more effective pharmacologic approaches are needed.

Drugs that act by inhibition of ion channels remain the main strategy to terminate AF, especially multi-ion channel blockers in comparison to selective-ion channel blockers. Blockade of atrial-selective sodium ion (Na^+) channels may effectively and safely suppress AF and concurrent potassium ion (K^+) channels inhibition may increment efficacy. Selectively targeting atrial K^+ channels may also be potentially relevant [61].

As for ablation, there are two different approaches, catheter ablation and surgical ablation. The former usually targets the pulmonary vein [62] with its complete electrical isolation as a primary goal. Although isolation is sufficient to suppress paroxysmal AF in most patients during one year of follow-up, in most cases of persistent AF substrate modification is needed [13]. Catheter ablation shows better results in maintaining SR than antiarrhythmic drugs but recurrences do occur ([63] cited by [13]). Besides, this procedure presents limitations in the reconnection of the isolated veins and it may lead to iatrogenic atrial tachycardia [13]. Surgical ablation, on its turn, creates transmural lesions capable of interrupting macroreentrant circuits that take part in sustaining atrial flutter or fibrillation. Studies show that rehabilitation is achieved in 75% to 95% of cases. Anyhow, this procedure is still hardly used given its complexity and high likelihood for major complications [54,64].

As previously stated, the rising prevalence, the complications associated with AF and its mortality rates pose a major threat to public health. On that account, identifying individuals who are at risk of developing the disease and, thus, may benefit from primary prevention has a considerable significance nowadays. In addition, the ability to predict AF within few months of onset may concede management approaches to have greater impact on outcomes. Nevertheless, no screening test has been developed to predict new cases of AF in at-risk patients [13].

Predicting AF's progression is also an interesting and important feat, but there is only one available tool to do so. According to the HATCH score [Heart failure, Age, previous Transient ischaemic attack or stroke, Chronic obstructive pulmonary disease and Hypertension (one point for each)], 50% of the subjects with scores larger than five progress to persistent AF in juxtaposition to the 6% of patients with a score of zero that do so [65]. Still, this score is in need of validation regarding assessment of differential weighing of components like systolic versus diastolic heart failure, hypertension with or without left ventricular hypertrophy and additional risk factors [13].

From the prevailing AF's clinical paradigm, it is apparent that the current diagnostic and prognostic methods lack specificity in differentiating AF from other medical conditions. On that account, new methods are required either to be used alone or in conjunction with the current approaches. The measurement of molecular disease markers might potentially be an effective screening and prognosis strategy, given that changes in the expression of such markers might be specific to the target disease. Therefore, the challenge is to identify these potential disease markers, in order to improve the screening and prognosis' efficacy of AF. Additionally, the same markers might represent themselves possible therapeutic targets.

2.4 Mechanisms Underlying the Pathophysiology of AF

AF is not only a result of ageing, it is also an expression of myocardial damage caused by modifiable and nonmodifiable risk factors. Even though AF is associated with many clinical conditions, the mechanisms underlying these associations are not completely understood [13]. Although AF was first identified in 1909, the notion that AF tends to propagate itself was only discovered in 1995 by Wijffels and colleagues [66].

Such process is likely a result of electrophysiological remodelling, probably related to the recycling of ion channels. This electric remodelling is marked by a reduction in the L-type calcium ion (Ca^{2+}) current, which leads to short-lived action potentials and the loss of the ability to adapt the heart rate to the duration of the action potential [67]. Electric remodelling is also implicated in the decrease of atrial contractility observed in AF and, as such, both events occur alongside [68].

In terms of structural changes, modifications like redistribution of nuclear chromatin, loss of myofibrils, accumulation of glycogen, alterations in mitochondrial shape and size, fragmentation of the sarcoplasmic reticulum (SPR), Z-line disruptions, and complete interruptions of myofibrils have been identified in animal models [69,70]. Post-mortem analysis of human samples have shown the presence of myocardial inflammation and fibrosis confined to the atrial myocardium, but not present in ventricular walls [71]. In fact, differentiation of fibroblasts into myofibroblasts is observed in AF.

Such cells exert a paracrine activity on cardiomyocytes, vital to electrophysiological and structural remodelling [72]. Structural remodelling, however, takes place on a longer timescale and it is likely associated with age, hypertension and multiple comorbid cardiac diseases [12]. It is known that both types of remodelling can be induced by heart failure [73,74] and atrial ischemia [75]. The mechanisms by which ageing induces AF are poorly understood, but anisotropy due to myocardial fibrosis [76] and connexin redistribution [77] are likely involved.

AF's pathophysiology comprises three major stages, initiation, maintenance and progression [12,73]. Interestingly, not only is AF generally initiated and maintained in the left atrium, the cycle length is also, in most cases, shorter in the left than in the right atrium [15]. In fact, areas with shorter cycle length are thought to be the critical substrate for driving or maintaining the fibrillatory circuits in AF [11].

In physiological conditions, atrial cells stand at a negative intracellular membrane potential, called the resting potential, become very positive when depolarized and go through a series of repolarizing states, including a plateau phase, to enter the resting potential phase once again [73]. Atrial action potentials start with the activation of voltage-dependent Na^+ channels, leading to cell depolarization {[72], **Figure 1 – (1)**}. During the action potential plateau, Ca^{2+} enters cardiomyocytes through voltage-dependent channels [**Figure 1 – (2)**], triggering Ca^{2+} release from the SPR through ryanodine receptors [RyRs, **Figure 1 – (3)**]. This systolic Ca^{2+} release is responsible for cardiac contraction [73]. Time-dependent delayed-rectifier K^+ currents and the transient-outward K^+ current allow cell repolarization and control the action potential's duration [**Figure 1 – (4)**]. The basal and acetylcholine-dependent inward rectifier K^+ currents control final repolarization and determine resting membrane potential [**Figure 1 – (5)**]. During diastole, Ca^{2+} is handled by the electrogenic Ca^{2+}/Na^+ exchanger (NCX), which transports three Na^+ into the cell and one Ca^{2+} outwards, resulting in a depolarizing inward current [**Figure 1 – (6)**]. Ca^{2+} is also removed from the cytosol into the SPR via the SPR Ca^{2+} ATPase (SERCA) pump [**Figure 1 – (7)**]. These processes restore low Ca^{2+} concentrations characteristic of the resting state and warrant atrial relaxation during diastole [72].

The beginning and extent of an AF episode relies on several electrical/structural triggers and substrates. The substrate sustaining AF comprises altered electrophysiological properties and altered structural properties of the atrium [11]. Although triggers are diverse (e.g.: sympathetic or parasympathetic stimulation, bradycardia, atrial premature beats or tachycardia, accessory atrioventricular pathways and acute atrial stretch), they do not cause the arrhythmia in the absence of other contributors.

Ectopic foci occurring in sleeves of atrial tissue within the pulmonary veins or vena cava junctions also constitute AF's triggers ([78] cited by [45]). Ectopic activity and re-entry are major mechanisms responsible for AF (**Figure 2**). Focal ectopic activity is probably caused by delayed afterdepolarizations (DADs) and early afterdepolarizations

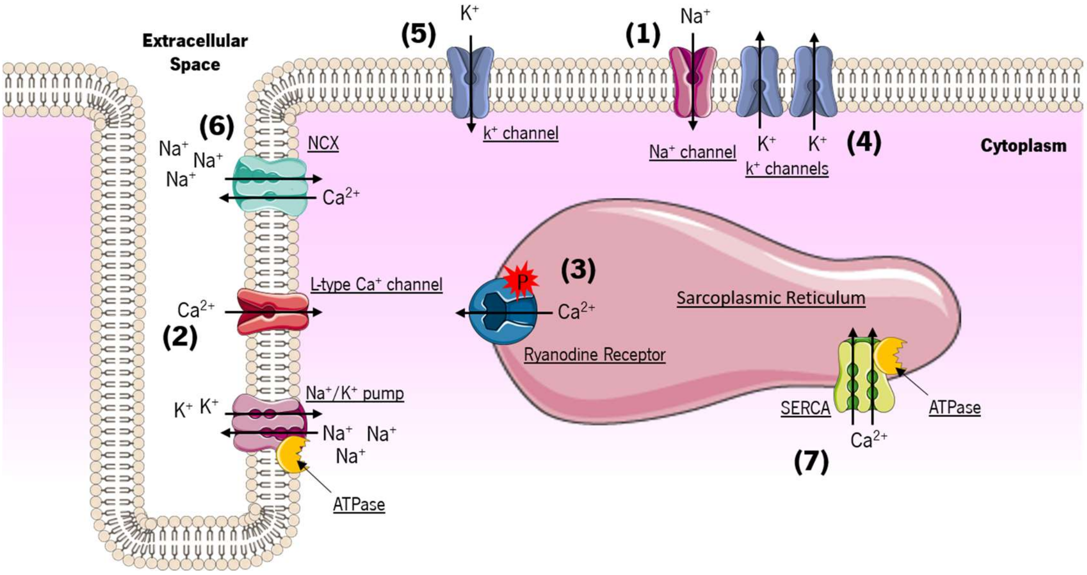


Figure 1 – Phases of the cardiac action potential **(1)** Na$^+$ channels are activated resulting in an influx of Na$^+$ to the cell (depolarization); **(2)** L-type calcium channels are activated allowing Ca^{2+} to enter the cell, which leads to Ca^{2+} release from the SR [**(3)**, plateau]; **(4)** The K$^+$ outflow from the cell turns the plasmatic membrane more negative (repolarization); **(5)** inward rectifier K$^+$ currents control final repolarization; **(6)** NCX transports 3Na$^+$ into the cell and 1Ca^{2+} out of the cell; **(7)** Ca^{2+} is removed from the cytoplasm into the SR through SERCA pump, resulting in the resting state. **Abbreviations:** NCX, Ca^{2+}/Na+ exchanger; SERCA, SPR Ca^{2+} ATPase; SR, Sarcoplasmic Reticulum.

(EADs). DADs [**Figure 2 – (1)**] result from abnormal and diastolic leak of Ca^{2+} through the RyRs and are promoted by an increased load of Ca^{2+} into the SPR and dysfunction of the receptors [72]. RyRs' function is modulated by phosphorylation and hyperphosphorylation causes them to become leaky and, thus, arrhythmogenic [79]. Excess of Ca^{2+} activates the NCX, producing a depolarization current. If the DADs are large enough to reach threshold, an ectopic action potential is triggered [72]. EADs [**Figure 2 – (2)**] are generally associated with prolonged duration of action potentials, due to the loss of repolarizing K$^+$ currents [80] or an excessive late component of noninactivating Na$^+$ current [81]. During a normal action potential, L-type Ca^{2+} channels undergo voltage and Ca^{2+}-dependent inactivation. However, prolonged action potential duration allows these channels to recuperate from inactivation, resulting in an inward current of Ca^{2+} and, consequently, an EAD [72].

Re-entry requires appropriate tissue properties that can either be caused by altered electrical properties or by fixed structural changes [72]. It can occur around an anatomical obstacle when each point in the pathway is able to regain excitability before the next impulse arrives. The possibility of anatomical re-entry is controlled by the wavelength [83]. Re-entry can also originate when premature impulses conduct unidirectionally

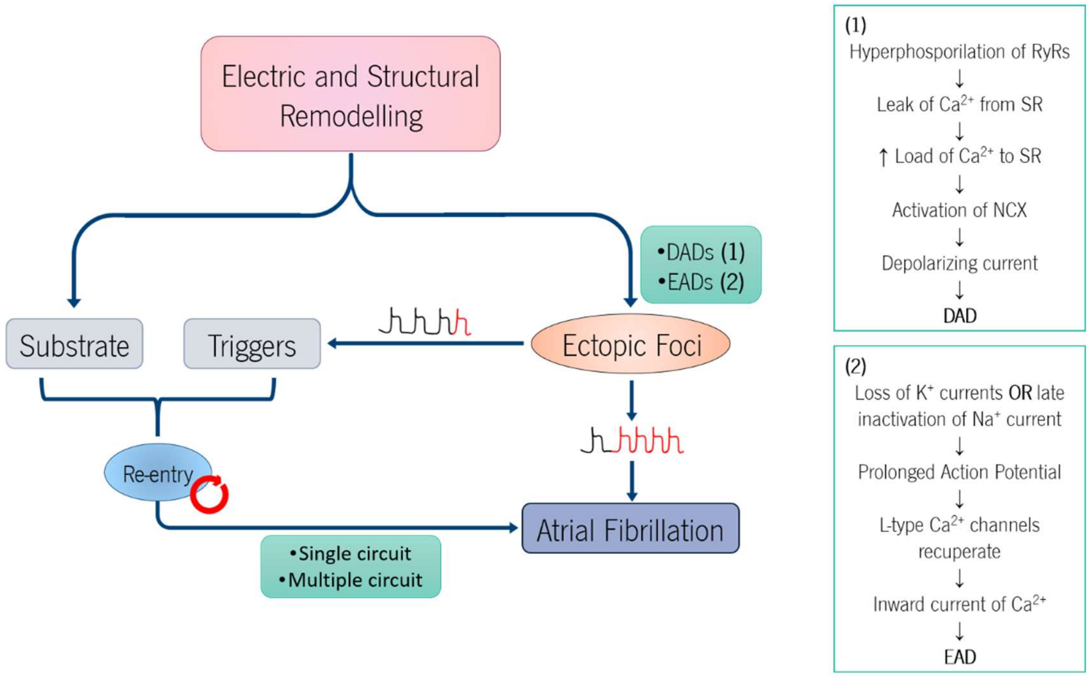


Figure 2 – AF's persistence mechanisms. Electric and structural remodelling favour the development of ectopic foci, more specifically through the cause of DADs and EADs. **(1)** Hyperphosphorylation of RyRs leads to Ca^{2+} leakage from the SR, which is promoted by an increase in the load of Ca^{2+} to the SR. The excessive amounts of Ca^{2+} activate the NCX, resulting in a depolarizing current and, consequently, an EAD. **(2)** The loss of K^+ currents or a late inactivation of the Na^+ current lead to a prolonged action potential, which allows the L-type Ca^{2+} channels to recuperate. The inward current of Ca^{2+} results in an EAD. The resulting ectopic foci causes AF and other triggers. Such triggers, together with sustaining substrates caused by structural and electric remodelling, create single or multiple re-entry circuits that perpetuate AF. **Abbreviations:** AF, atrial fibrillation; DADs, delayed afterdepolarizations; EADs, early afterdepolarizations; NCX, Ca^{2+}/Na+ exchanger; RyRs, ryanodine receptors; SR, sarcoplasmic reticulum. Adapted from [82].

around an initial refractory border [72]. Furthermore, slowed conduction is also able to lead to re-entry since slower conduction of an impulse leaves additional time for refractoriness to dissipate [73]. Together, atrial dilation and fibrosis create longer potential conduction pathways, slower conduction and impose conduction barriers that favour both initiation and maintenance of multiple re-entry circuits that sustain AF [84]. AF-related re-entry might occur as a single circuit, involving one primary re-entry circuit driver, or as a multiple circuit, involving several simultaneous dyssynchronous re-entry circuits. Functional re-entry is promoted by atrial tachycardia remodelling (ATR) of atrial electrical properties, which in turn is caused by the very rapid activation produced by AF. ATR leads to spatial heterogeneity that generates multiple circuit re-entry [8]. Mitral valve disease, which induces atrial dilation, conduction disturbances and electrical remodelling, also favours re-entry [85,86].

2.5 Biomarkers in Atrial Fibrillation

The International Programme on Chemical Safety, led by the World Health Organization in coordination with the United Nations and the International Labor Organization, has defined a biomarker as “any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease” ([87] cited by [88]). Biomarkers can be anything from pulse and blood pressure to measurements of blood and other tissues [88]. Biomarkers can, for instance, be proteins identified and measured through proteomic and other ancillary techniques, showing alterations in a certain disease, thus, carrying valuable diagnostic or prognostic information, such as allowing the determination of the likelihood of its recurrence [76].

Additionally, such proteins or their peptide products can themselves be novel molecular targets for drug design [15]. Furthermore, from a clinical perspective, they assist physicians in the management of a patient status, being indicative of different stages in the development of a disease. In some cases, they can even be detected prior to the befall of a disease, reinforcing their role in disease prevention. In summary, they can be useful for diagnosis, prognosis and therapy monitoring [89]. Besides, it should be noted that applied strategies for biomarker discovery can even be of great interest to shed light into the pathophysiological mechanisms of the disease [15].

Biomarkers have been widely used in the diagnosis and management of myocardial infarction and heart failure but not in AF [90–92]. Only recently, did they start coming into view as promising predictors of AF risk [93]. In 2016, the European Society of Cardiology guidelines on AF recommended the use of biomarkers such as high-sensitive troponin and natriuretic peptides to further refine stroke and bleeding risk in AF patients [4]. These guidelines, however, recommend the use of biomarkers for an evaluation of the risk of complications resulting from AF and not for diagnostic or prognostic purposes of AF itself.

Anyhow, a growing number of studies, either by proteomic approaches per se or by the use of other techniques to study specific proteins, have been associating AF with several protein alterations and consequently uncovering surrogate biomarkers. Most of these, however, focus on one or two proteins or on a selected group of proteins, using techniques such as One(1)-Dimensional SDS-Polyacrylamide Gel Electrophoresis (1D SDS-PAGE), Two(2)-Dimensional gel Electrophoresis (2DE), Enzyme-Linked Immunosorbent Assay (ELISA), western blot and other immunoassays. ELISA, for instance, is used in a variety of studies to detect changes in protein's expression levels. Fu *et al.* (2011) [94] quantified the plasma levels of von Willebrand factor and p-selectin of AF patients using ELISA. P-selectin levels were higher in AF patients compared to controls, but there were no statistically significant differences regarding von Willebrand factor. Gordon *et al.* (2016) [95] also used ELISA to quantify the levels of galectin-3 and fibroblast growth factor 23. In this study, there was no statistically significant differences between the AF group and the control group nor between patients with and without AF recurrence.

Troponin, for instance, is one of the most frequently studied proteins and is hoped to aid in the management of AF, particularly new-onset and postoperative AF, myocardial infarction in AF and prognosis of AF [96]. Elevated troponin levels have been associated with increased incidence of AF [97–99], but the optimal cut-off to determine the risk of AF is still unclear [97,98]. Therefore, the use of troponin screening to predict the risk of AF incidence is alluring but premature [96]. As for the risk of postoperative AF, preoperative values of troponin were not associated with risk of postoperative AF [100]. Additionally, one study reported an association with postoperative cardiac-troponin T levels [101] and a second study with postoperative cardiac-troponin I levels [102]. However, a third study was unable to validate the second finding [103]. Comparably to new-onset AF, there is also no cut-off value to determine the risk of postoperative AF [96]. In order to distinguish between myocardial infarction and AF, attempts to ascertain an optimal cut-off for troponin levels have been made [104,105]. Persistently detectable troponin values separated by 3 months indicated worse prognosis compared with patients with undetectable or transiently detectable levels [106]. In brief, more studies are necessary to evaluate the power of troponin in detecting the risk of AF, especially due to contradictory results, requiring consensus. Nonetheless, troponin remains a potential biomarker for diagnosing and prognosticating AF.

Alike troponin, Brain natriuretic peptide (BNP) and N-terminal proBNP (NTproBNP) have been associated with the development of AF, either postoperatively or not. They have also been associated with diagnosis of heart failure in AF and have been proposed as predictors of the success of current direct cardioversion for AF [96]. However, the threshold levels of BNP and NTproBNP for the diagnosis of AF are yet to be determined [107]. As an example, in 2007, Matsuura *et al.* [108] investigated the relationship between the plasma BNP levels and the occurrence of AF in nonobstructive hypertrophic cardiomyopathy. Their results showed significantly higher plasma BNP levels in the paroxysmal AF group and in the long-standing persistent AF group than in the SR group. As a result, the authors concluded that plasma BNP levels are clinically useful for identification of nonobstructive hypertrophic cardiomyopathy patients who are at risk of developing AF. BNP and NTproBNP have also been asserted as independent predictors of AF in the Cardiovascular Health Study [109]. C-reactive protein (CRP) is yet another protein with biomarker potential for AF over a median follow-up of 7.8 years in the same study [110].

Despite proteomic studies per se are currently not as voluminous for AF as studies focusing on one protein or a particular set of proteins, they are extremely relevant. Proteomics is the large-scale study of proteins in a complex biological sample (biofluids, cells, tissues, etc) at a given time [82], using high-throughput techniques, such as mass spectrometry (MS), to identify and detect changes in multiple proteins, unveiling several potential biomarkers. Hence, proteomics' tools can be of extreme importance in the process of discovery and identification of potential biomarkers and in the creation of a biomarker panel for a target disease. Likewise, it has the potential to aid in the implementation of novel and standardized diagnostic and prognostic approaches for AF and other conditions.

In 2009, Mondrego *et al.* [111] undertook a proteomics approach to evaluate the expression of proteins associated with the cytoskeleton, energetic metabolism, and cardiac cytoprotection in left atrial appendages (LAA) and right atrial appendages (RAA), obtained from patients with mitral valve disease both in SR and in permanent AF. Proteins were separated with 2DE and identified by MS and more than 30 proteins were analysed. Cardiac α-actin isotypes 1 and 2, tropomyosin α- and β-chains, and myosin light chain embryonic muscle/atrial isoform were found to be overexpressed in the LAA of AF patients compared to the LAA of SR patients. Different cytoskeleton-associated protein levels measured from RAA samples, as well as different energetic metabolism-associated proteins levels measured from both LAA and RAA samples were elevated with respect to those from SR patients. The expression of proteins associated with cardiac cytoprotection, such as gluthatione-S-transferase, heat shock protein (HSP27), and different 60 kDa heat shock protein (HSP60) isotypes, were higher in the RAA of AF patients compared to the RAA of SR patients. This study is a perfect example of how MS-based proteomics experiments can detect differentially expressed proteins, with potential biomarker value.

Thousands of claimed biomarkers are reported in thousands of biomedical papers [109]. Such a tremendous amount of data led to the belief that effective biomarker-based diagnosis would rapidly unfold [112]. However, even though several molecules were and are being investigated as candidate biomarkers, most have not been validated for routine clinical practice [113]. Indeed, the vast majority of biomarkers lack the validation or downstream development and refinement necessary for clinical translation [112].

This comes to show the importance of extracting and re-evaluating information on disease-related surrogate biomarkers from scientific publications. In fact, curation of relevant literature might aid in the creation of a biomarker panel for a specific disease. However, establishing correlations between diseases and changes in biomarkers is not enough; the amount of associations that have already been established, along with the ones that are yet to be uncovered, and the fact that changes in a certain biomarker may not be specific to a disease, lead to the need of a biomarker ranking, which can be achieved by scoring the true biomarker value of biomolecules, such as proteins, through appropriate functions, weighing several parameters. Such score could weigh, among others, the number of evidences supporting the association, the existence of independent validations either intra- or inter-studies, the performance of a quantitative analysis, among others. From this point forward, this dissertation will only focus on protein biomarkers; the term “biomarker” will refer to protein biomarker for simplification.

2.6 How Text-mining and Bioinformatics Resources May Help Defining a Biomarker Panel for AF

Multiple areas of research from molecular biology to machine learning are partnered to better understand complex biological systems, such as cells, tissues and the human body [114]. Bioinformatics is now a key field in this regard, because it allows to deal with the tremendous volume of publicly available data ([115] cited by [116]). Proteomics, for instance, contributes with massive MS datasets, which represent tremendous amounts of information, the so-called big data. Consequently, bioinformatics became fundamental and indispensable for life sciences [117]. On a different note, it is also highly dependent on it, given that laboratory work is responsible for the translation of matter of life into data, which can then be assessed by bioinformatics [118].

There has been an increase in the number of diagnostic breakthroughs and in successful efforts in identifying patients more susceptible to certain diseases or that will maximally benefit from certain treatments using biomarkers. Although oncology was the main area of research in this field, there have been considerable advances in other areas, for instance respiratory, infectious and inflammatory diseases, in the last ten years [119]. These advances are reflected in the vast amount of literature available from online databases, such as PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), regarding biomarker-disease associations. Effectively extracting such associations from papers will potentially enable the discovery and development of new therapeutic targets and patient segment biomarkers [119]. There are several online databases built and maintained for this purpose, capable of gathering in one place valuable information, such as The Human Metabolome Database (HMDB; http://www.hmdb.ca/diseases), Global Online Biomarker Database (GOBIOM; https://gobiomdb.com/login.jsp) and DisGeNET (http://www.disgenet.org/). HMDB is an online database that reports information on several associations. However, it only covers metabolites [120]. GOBIOM is yet another database that provides information on several types of markers for multiple conditions, with its reported utilities like diagnosis, prognosis, monitoring disease progression, among others. Information is gathered by over 200,000 sources, including clinical trials, scientific conferences, regulatory-approved documents, literature databases, patents, etc [121]. DisGeNET stores a large amount of gene/protein-disease associations and will be discussed further ahead. Despite the ability to perform a quick search for a given condition/marker, these online repositories may not comprehend the entire collection of biomarkers for a condition. This is due to the ever-growing number of biomedical papers and the requirement of a careful survey of the same [122]. Hence, the process of text-mining, either manual or with the aid of bioinformatic tools gains relevance in this field. The choice of which curation solution works best for a given researcher, laboratory or organization varies [123], according to the number of relevant retrieved articles and the amount of information to extract. However, manual and automated text-mining have different advantages, which will be debated in the following sub-section.

2.6.1 Manual vs Automated Text-mining

Although, as previously mentioned, there are already multiple online repositories of biological data, most of the relevant information is still maintained in textual format [124]. Likewise, the curation of literature represents a crucial aspect towards the annotation of important information. The first step of the curation process is the gathering of relevant research papers, which often involves a paper-by-paper review by the curators [125]. However, automated forms of information retrieval (IR) have been developed, which help to prioritize literature curation, such as @Note [126]. The final collection may range from 100 to 1000 papers a month [125]. The next step is the actual curation of each of the selected articles to extract pertinent information. The type and amount of data to be extracted is highly dependent on the ultimate goal of the curation process. Even though many experts are involved in this process, manual data curation is time-consuming [122]. Also, manual curation can be biased by limiting journals and articles due to resource restrictions and journal value [123]. Nonetheless, the resulting datasets are of high quality and relevance [127], due to an overall accuracy of 90% of expert curators [123].

With the discussed issues in view, biomedical text-mining emerged as a new research field [128]. This automated form of curation encompasses IR, information extraction (IE) and Hypothesis generation [126]. Hypothesis generation tries to conciliate data from experimental procedures or in silico experiments with annotations derived from the literature [129]. Initially, such computational solutions could not compete with the accuracy and completeness of the gold standard manual curation. Nowadays, that is no longer the case; the technological advances allowed the improvement of automated curation. Automated curation systems scan and retrieve papers without associated bias and are only limited by legal issues and licensing fees. Plus, new ontologies can easily and rapidly be added to include new terms and concepts in biology [123]. Howbeit, higher error rates and less overall relevance are still a reality [127].

2.6.2 Ranking Biomarkers

The curation of relevant literature might aid in the creation of a biomarker panel for a specific disease. However, establishing correlations between diseases and changes in biomarkers is not enough. After extracting gene/protein expression data or molecule-disease associations from the literature or from databases, the next challenge is to define the relevance of the molecules towards their biomarker potential. The amount of associations that have already been established, along with the ones that are yet to be uncovered, and the fact that changes in a certain biomarker may not be specific to a disease, lead to the need of a biomarker ranking, which can be achieved by scoring the relative biomarker value of biomolecules through appropriate functions with high discriminatory abilities, weighing several parameters. Only after this step, may a group of molecules proceed in the pipeline towards biomarker validation and implementation. In that regard, several scores can already be found in the literature and are described in the next subsections and summarized in **Table 2** and **Table 3**. These scores can be based on several criteria, already mentioned in section 2.5, comprising, therefore, a more objective and independent approach to define the biomarker panel.

2.6.2.1 Scoring Approaches to Rank Associations Extracted by Text-mining

Many scoring approaches based on text-mining are focused on the number of papers studying an association. Consequently, the higher the number of studies reporting the association, the higher the score will be. Notwithstanding, these often ignore the variation of the expression levels/quantification values of the selected markers. Xu *et al.* (2016)[119] proposed four different methods (**Table 2**) to rank disease-related genes based on co-occurrence frequency, paper citations and author information. The first method simply ranks different marker-disease pairs, where the disease is the same, by the number of distinct journal articles in which they co-occur. The second method is based on the PageRank algorithm which was proposed by Larry Paige and Sergey Brin in 1998 [130] and operates on the idea that the more important a website is, the more websites will link to it. Therefore, the importance of a website is based on the number and significance of the websites connected to it. In the study by Xu *et al.* (2016)[119], the PageRank of an article is higher the more articles cite it and the more influential those articles are. The third ranking function (suppressed PageRank method) is improved by consideration of the authors. Considering that researchers who focus on specific diseases or genes may write about the same gene-disease pair in multiple publications, it makes sense that the contribution of the multiplicated evidence should be somehow corrected. However, while correcting for some data homogeneity, this type of scoring drags another issue, related to the rationale and experimental approach of the studies taken by the same authors. In fact, there may be the case where the same marker-disease association is described by the same author(s), but in independent study populations, distinct biological sample types and even with different experimental techniques, sometimes used for inter-validation. In such cases, a correcting factor should be minded. Finally, the last method (time-weighted PageRank method) uses a PageRank function adjusted by a time factor. This adjustment is placed because of the assumption that recently published articles may have "less exposure" for citation and have, consequently, fewer citations than those published before them. This last method also suppresses gene-disease pairs mentioned multiple times by the same authors. In order to evaluate which of the four methods performed better, ten diseases were extracted from DisGeNET and the mean reciprocal rank of ranks obtained was computed for each method. Overall, the Suppressed PageRank method achieved better results compared to the other three methods.

Bravo and colleagues' (2013)[113] score is obtained as the product between the inverse document frequency (IDF) of the association and the normalized frequency of the association (NF). The IDF consists of the logarithm of the total number of abstracts considered in the study, which correspond to a pool of articles extracted from a specific query to PubMed, divided by the number of abstracts containing the association in question. The function NF is computed as the quotient between the number of times an association is found in a certain abstract and the maximum frequency of any association in that same abstract. This approach is based on the premise that if a biomarker and a disease are mentioned together in the same sentence, then there is a high probability of them being associated to each other. Withal, this is not always verified, because results might not be statistically significant. Moreover, whilst the analysis time may be reduced when considering only the abstract, all the information in the body of the article is neglected, which could have given indication of other pertinent associations.

DisGeNET (http://www.disgenet.org/web/DisGeNET/menu/home;jsessionid=17dgwkd84j64a14noky9ny3epk), which might be the most popular online repository for marker-disease associations, integrates human gene-disease associations from various expert curated databases and text-mining derived associations. It also presents approaches to rank the gene-disease associations (GDA) and the variant-disease associations (VDA), according to their level of evidence (**Table 2**). In both methods scores range from zero to one, and consider the number and type of sources and the number of publications supporting the association [131]. The GDA system also contemplates the number of different models (mouse, rat, human, etc) in which the association was studied. The higher the number of models, the higher should the strength of the study be and, consequently, the higher the score.

The approaches described by Xu et al. (2016)[119] and Bravo et al. (2013)[132] and DisGeNET's scoring systems take into account the number of publications supporting an association. However, these do not consider the type of variation observed when the biomarker-disease association was conveyed. Simply knowing that a certain biological parameter, such as a gene or its corresponding protein/peptide or related metabolites, is associated with a disease does not give us any information with respect to the specific variation of such marker. In other words, there is no knowledge whether the entity's levels are higher or lower than the recommended values in the affected patients. Furthermore, they also do not consider the coherence between findings. DisGeNET does present an Evidence Index, which indicates the existence of contradictory results in publications regarding positive and negative associations, but this index does not consider contradictions in terms of alterations in opposite directions.

2.6.2.2 Ranking Approaches Based on Expression Data

The previous approaches do not consider the use of expression data; still, acknowledging the degree of variation and its direction is important to achieve a more accurate and quantitative biomarker panel. Ernst et al. (2017)[132], proposed an heuristic approach (**Table 3**) to rank GDAs based on expression levels and on a gene interaction/regulation network. The FocusHeuristics algorithm computes three scores: the log fold change (LFC), i.e. the log-transformed difference of gene expression between two conditions, the differential link score (LSd), which is the sum (for activation or unspecified links) or the difference (for inhibition) of the LFCs of the connected genes, and the interaction link score (LSi), which is the lowest value of the sum of the expression levels of the connected genes for each condition, representing the activity of a link in the graph for both conditions. This algorithm generates a new network, keeping the nodes of the reference network that pass at least one of the thresholds set for each score. Although it may be true and logic that when the expression of a gene is altered there are downstream effects, the premise on which FocusHeuristic is based, this is not always verified, which means that the final condensed network may include genes whose expression is not altered and therefore does not take part in the pathophysiology of a certain disease. This happens because the activation or inhibition of a certain gene is not directly controlled by a second gene per se, but by the product of that second gene, whose expression may not be altered due to regulation steps that take part after the transcription process.

A similar approach (**Table 3**) was proposed by Yu *et al.* (2015)[133], where both the expression levels and an interaction network are considered, to identify risk genes associated with myocardial infarction. In this case, the suggested system takes into consideration the fold change value for the expression level of the node and the nodes connected to the selected node. With this scoring system, called neighbourhood scoring algorithm, the influence of the “diseased” genes on their connected genes is inferred; if the score is >0, the node and its connected nodes are highly expressed, and if the score is <0, the expression of the nodes is low.

Although the variation is weighed in both methods, no attempts to understand its direction are made. Additionally, the frequency of reports describing a particular marker-disease association is not weighted as well as the coherence between findings. However, the usage of an interaction network, in both approaches, means that additional GDAs could be identified based on gene-gene regulation. Also, quantitative data regarding the variation of the expression levels/quantitative values is considered.

In spite the fact that a few scoring functions were already developed by other authors to rank genes/proteins according to their biomarker potential, as previously reported, several limitations can be found in those methods. Thus, after taking an extensive literature search with relaxed queries and mining a massive volume of reports, protein data was extracted and scored to define potential biomarker panels for AF diagnosis or prognosis. In this sense, a mathematical scoring function was developed to weigh the importance of proteins, based on several parameters (consensus of studies supporting the association, median fold-change of a protein and number of diseased individuals in the study), and to minimize the limitations of the currently known approaches.

Table 2 – Existing ranking scores for associations extracted by text-mining, respective potentials and limitations.

Name	Formula	Potentials	Limitations	Reference /Database
Frequency-based	?	• Co-occurrence frequency in the literature is considered to identify disease-related genes.	• Categorical information regarding the type of variation is ignored; • Quantitative data regarding the variation of the expression levels is not considered.	Xu *et al.* (2016)
PageRank-based	$S(g,d) = \sum_{a \in C_{(g,d)}} pr(a)$ **Where:** g → gene; d → disease; $C_{(g,d)}$ → set of all the articles that contain the (g, d)pair; $pr(a)$→ PageRank of paper a.	• Information is prioritized according to the degree of data accession (it assumes that most cited articles are more important).	• Categorical information regarding the type of variation is ignored; • Quantitative data regarding the variation of the expression levels is not considered; • Most recent papers and, thus, recently reported marker-disease associations have less exposure and may be neglected.	Xu *et al.* (2016)
Suppressed PageRank	$S = \sum_{a \in C_{(g,d)}} w_a(g,d) \times pr(a)$ **Where:** $w_a(g,d) = \frac{\sum x \in l \frac{1}{\lvert C_x \rvert}}{\lvert l \rvert}$ **Where:** l → author list of paper a; C_x → number of papers author x wrote about (g, d); g → gene; d → disease; $C_{(g,d)}$ → set of all the articles that contain the (g, d) pair; $pr(a)$ → PageRank of paper a.	• Information is prioritized according to the degree of data accession It tries to suppress repeated contributions by the same authors.	• Categorical information regarding the type of variation is ignored; • Quantitative data regarding the variation of the expression levels is not considered; • Most recent papers and, thus, recently reported marker-disease associations have less exposure and may be neglected; • While suppressing repeated conclusions by the same authors, it ignores if different reports from the same author entails different populations, biological samples and technical procedures.	Xu *et al.* (2016)

Table 2 – Existing ranking scores for associations extracted by text-mining, respective potentials and limitations (continued).

Name	Formula	Potentials	Limitations	Reference /Database
Time-weighted PageRank	$pr(u) = d \sum_{v\in B(u)} \frac{pr(v)}{N_v} + (1-d) \times T_u$ **Where:** d → disease; T_u → time factor related to each paper's year of publication.	- Information is prioritized according to the degree of data accession; - The relevance of each article is balanced by the time of publication.	- Categorical information regarding the type of variation is ignored; - Quantitative data regarding the variation of the expression levels is not considered; - Repeated contributions of the same authors for a given marker-disease relationship are neglected	Xu *et al.* (2016)
Variant of the Inverse Document Frequency model	$S_{DB} = idf(DB,A) \times \sum_{i=1}^{\vert A\vert} \alpha f(DB,A_i)$ **Where:** $idf(DB,A) = \log_{10} \frac{\vert A\vert}{\vert\{\alpha \in A: DB \in a\}\vert'}$ $af(DB,A_i) = \frac{f(DB,A_i)}{\max\{f(xy,A_i): XY \in A_i\}'}$ D → disease; B → biomarker; $\vert A\vert$ → total number of abstracts; A_i → the ith abstract; $f(DB,A_i)$ → number of times the association between D and B occurs in A_i.	- Associations are ranked according to the frequency of the reports describing them; - The association between the marker-disease pair at scope is weighted according to the maximum number of associations described in the abstract for any other pair.	- Categorical information regarding the type of variation is ignored; - Quantitative data regarding the variation of the expression levels is not considered; - Repeated contributions of the same authors for a given marker-disease relationship are neglected; - Analysis is restricted to the content of the abstract.	Bravo *et al.* (2013)

Table 2 – Existing ranking scores for associations extracted by text-mining, respective potentials and limitations (continued).

Name	Formula	Potentials	Limitations	Reference /Database
The GDA Score	$S = C + M + \sum_{k=1}^{3} L_k$ **Where:** $C = \begin{cases} 0{,}6\ if\ N_{sources_i} > 2 \\ 0{,}4\ if\ N_{sources_i} = 2 \\ 0{,}2\ if\ N_{sources_i} = 1 \\ 0\ otherwise \end{cases}$ **Where:** $N_{sources_i}$ → number of curated sources supporting a GDA; i ∈ UNIPROT, CTD, PSYGENET, ORPHANET, HPO. $M = \begin{cases} 0{,}16\ if\ N_{models} = 2 \\ 0{,}08\ if\ N_{models} = 1 \\ 0\ otherwise \end{cases}$ **Where:** N_{models} → number of animal models for a GDA; Models ∈ Rat, Mouse from RGD, MGD, CTD. $L = \begin{cases} 0{,}08\ if\ \frac{N_{gd} \times 100}{N_{literature}} \geq 0{,}08 \\ \frac{N_{gd} \times 100}{N_{literature}}\ if\ \frac{N_{gd} \times 100}{N_{literature}} < 0{,}08 \end{cases}$ **Where:** N_{gd} → number of publication supporting a GDA in the source k; $N_{literature}$ → total number of publications in the source k; k ∈ GAD, LHGDN, BEFREE.	• Associations are ranked according to the frequency of the reports describing them (L); • The number of animal models supporting an association is duly accounted (M); • The curation level of the data source is taken into account (C).	• Categorical information regarding the type of variation is ignored; • Quantitative data regarding the variation of the expression levels is not considered; • No correction is attempted for marker-disease associations found by the same authors, in the same population, in the same biological sample or uncovered through the same technique.	DisGeNET

Table 3 – Existing ranking scores for expression data, respective potentials and limitations.

Name	Formula	Potentials	Limitations	Reference/Database
FocusHeuristics **- Differential Link Score (LSd)** **- Interaction Link Score (LSi)**	$LS_d = LS_d^{AB} = LFC^A + direction^{AB} \times LFC^B$ $\mathrm{LS_i} = \mathrm{LS_i^{AB}} = \min(\mathrm{E_t^A} + \mathrm{E_t^B}, \mathrm{E_c^A} + \mathrm{E_c^B})$ **Where:** A, B → nodes/genes; AB → edge/interaction between the two nodes/genes A and B; c, t → conditions; $\mathrm{E_C^{\mathit{G}}}$ → log expression level of gene G in condition C; LFC^A → log fold change of gene A; $direction^{AB}$ → direction of a directed edge: -1 for inhibitions, +1 otherwise.	• The type and magnitude of variation of a marker is considered; • Additional GDAs can be identified by considering the interactions between genes.	• The premise on which it is based is not always verified; • The frequency of reports describing an association is not considered; • The level of data curation is not acknowledged.	Ernst *et al.* (2017)
Neighbourhood Scoring	$S(i) = \frac{1}{2} \times FC(i) + \frac{1}{2} \times \frac{\sum_{n \in N(i)} FC(n)}{N(i)}$ **Where:** i → node; FC → fold change value for the expression level of the node; N(i) → number of the connection nodes to the selected node.	• The type and magnitude of variation of a marker is considered; • Additional GDAs can be identified by considering the interactions between genes.	• The frequency of reports describing an association is not considered; • The level of data curation is not acknowledged; • Indirect determination of the potential biomarker value.	Yang *et al.* (2017)

3. Methods

3.1 Literature Search

Independent PubMed queries were ensued up to 26th July 2017 to retrieve available proteomic and protein focused studies in AF, using the following keywords in separate queries: “atrial fibrillation AND proteomics” and “atrial fibrillation marker”. The literature search came up with 18 entries using the first set of keywords and 2255 entries with the latter, making a total of 2273 papers. A collection of potentially relevant articles (321 papers) was pre-selected after reading the titles and abstracts, whenever possible; if an abstract was not available but the title indicated a possibly pertinent study, the article was also retrieved. The following inclusion criteria were mandatory:

1) full-text English-written article;
2) publication in peer-reviewed journal;
3) proteomic or protein focused study;
4) study enrolling humans only;
5) study enrolling subjects with AF in comparison to healthy individuals, individuals who successfully cardioverted or individuals with other conditions except for other arrhythmias, and
6) study enrolling samples such as plasma, serum, whole blood, urine and atrial appendages.

As for exclusion criteria, studies which tried to establish associations based only on linear/logistic regressions or hazard/odd ratios, studies without control subjects and reviews and meta-analysis were left apart. In the end, 172 papers were included in the study and 149 were excluded.

3.2 Text-mining

The final assortment of publications was carefully analysed to extract data to Excel Spreadsheets, for further bioinformatic analysis. The subsequent fields were filled in: “Protein ID” (UNIPROT code), “Protein Name”, “Sample Type” (biofluid or tissue); “Sample” (plasma, urine, blood, left atrial appendages, right atrial appendages, pericardial fluid, serum), “Sample’s Treatment” {including: “Frozen” [yes (y) or no (n)], “Freezing Temperature”, “Centrifuged”, “Treatment”}, “Pathology Definition (Cases)” [including: “Age” (mean ± sd), “Disease Stage/Subtype” (paroxysmal, persistent, permanent, first-diagnosed, postoperative), “Recurrence study” (y or n), "Procedure", "Classification/Characterization" (additional medical conditions beyond AF), "n" (number of enrolled patients for discovery)], "Background Condition (Controls)" (including: "Age", "Procedure", "Classification/Characterization"), "Study Design", "Population Source", "Methodology", "MS approach", "Variation" [1 (overexpressed), -1 (underexpressed), 0 (unchanged) "N/A" (not applicable)], "Cases expression value", "Controls expression value", "Fold-change", "Fold-change formula", "p-value", "Statistical test", "DOI", "Article Title", "Year of publication", "Source of Data" (tables, figures, core text, etc, from which the expression values were retrieved). The fold-change was calculated as a ratio between the mean or median levels in diseased patients and the mean or median levels in control patients. Regarding the topic "variation", proteins were classified as overexpressed/underexpressed if levels of a certain protein were found higher/lower in patients compared to controls and the difference was statistically significant. A p-value of less than 0.05 was considered statistically significant; protein levels were considered unchanged if the p-value was higher. Entries were defined with "N/A" (not applicable) if there was no information in the paper whether differences were significant or not.

3.3 Scoring System

After analysing every retrieved article and extracting all the relevant information, a scoring approach was applied to subsets of the complete dataset. Four partial datasets were created according to the sample (atrial appendages, whole blood, plasma and serum) and each part was further subdivided into six subsets, whenever possible, in relation to the disease subtype ("All" - which includes first-diagnosed, paroxysmal, persistent, long-standing persistent and permanent AF as conditions -, paroxysmal AF, persistent AF, permanent AF, postoperative new-onset AF and postoperative AF recurrence), making a total of 24 subsets. The "All" subsets contain entries in which the AF state is not specified, entries which correspond to a mixture of at least two of the major types of AF (paroxysmal, persistent and permanent) and entries which correspond to one specific type of AF. The paroxysmal AF, persistent AF and permanent AF subsets, regardless of the sample, only contain entries which correspond to the type of AF specified in the subset's name. The postoperative new-onset AF subsets include entries in which the corresponding patients only developed AF after surgery and did not have history of AF. In all these cases, the control subjects had to be in SR and not have experienced AF in the past. The postoperative AF recurrence subsets include entries in which diseased subjects had AF prior to surgery and developed recurrence after surgery. Control subjects also had AF before surgery but did not suffer from recurrence. The scoring system was applied to each protein in each subset, in order to identify and rank proteins according to their biomarker potential.

$$S_d(p) = \overline{m}(FC_p)^x \times \frac{\sum C(E_{p,d})}{\sum C(E_p)}$$

Where:

d → direction of the variation;

p → protein;

$FC_{p,d}$ → Fold-change of p;

$$x \begin{cases} 1 \; if \; d = "up" \\ -1 \; if \; d = "down" \end{cases}$$

$$C(E_{p,d}) \begin{cases} 1 \; if \; n_{E_{p,d}} < \; \overline{m}(n) + IQR \\ 2 \; if \; n_{E_{p,d}} \geq \; \overline{m}(n) + IQR \end{cases}$$

Where:

$E_{p,d}$ → entry with p and in the direction d;

$C(E_{p,d})$ → contribution of $E_{p,d}$;

$n_{E_{p,d}}$ → number of enrolled patients in $E_{p,d}$.

n → number of enrolled patients in every entry of the dataset;

IQR → inter-quartile range.

$$C(E_p) \begin{cases} 1 \; if \; n_{E_p} < \; \overline{m}(n) + IQR \\ 2 \; if \; n_{E_p} \geq \; \overline{m}(n) + IQR \end{cases}$$

Where:

E_p → entry containing p;

$C(E_p)$ → contribution of E_p;

n_{E_p} → number of enrolled patients in

E_p.

Equation 1 – Simple scoring approach.

The formula (**Equation 1**), is applied in two cases, one that considers entries in which the protein was found to be overexpressed ($d = "up"$) and one which includes entries in which the protein was found to be underexpressed ($d = "down"$). The formula takes into consideration the median fold-change value and the degree of agreement between all the entries with the specific protein ($\boldsymbol{p}$), as a fraction of the number of entries supporting the association in the respective direction and the total number of entries with the protein.

The contribution of entries with a number of enrolled patients higher that the median plus the interquartile range (**IQR**) of enrolled patients in every entry is doubled.

Some proteins might not be scored in one or both directions because they might be defined as unchanged or "N/A" in every entry or not changed in the considered direction. In order for the protein to be considered a potential biomarker of the respective subset/condition, the computed score has to be higher than one (threshold).

The scoring function was implemented in form of an algorithm developed in Python 3.5. Five proteins whose score passed the threshold were chosen as the top potential biomarkers for the particular condition. The results concerning atrial appendages were ignored for the conditions "All", paroxysmal AF, persistent AF and permanent AF because the measurement of biomarkers' levels in such samples in the clinical practice is not viable, unless patients underwent cardiac surgery.

3.4 Bootstrap

In order to determine the uncertainty of the simple scoring approach, the method was applied in an iteration-based method, bootstrap, developed in Python 3.5. The bootstrap was performed for each subset but all entries corresponding to proteins which had a null score with the first scoring approach were eliminated. As such, each subset is divided in two parts, one which includes the entries regarding the proteins which were scored with the simple scoring approach in the direction "up" and one which includes the entries regarding the proteins which were scored in the opposite direction. Likewise, the input subsets are different according to the direction.

The bootstrap consists of creating partial datasets with the same size (**n_1**) as the input subset. Each entry has a probability (p) of being chosen for the new subset of size **n_2**

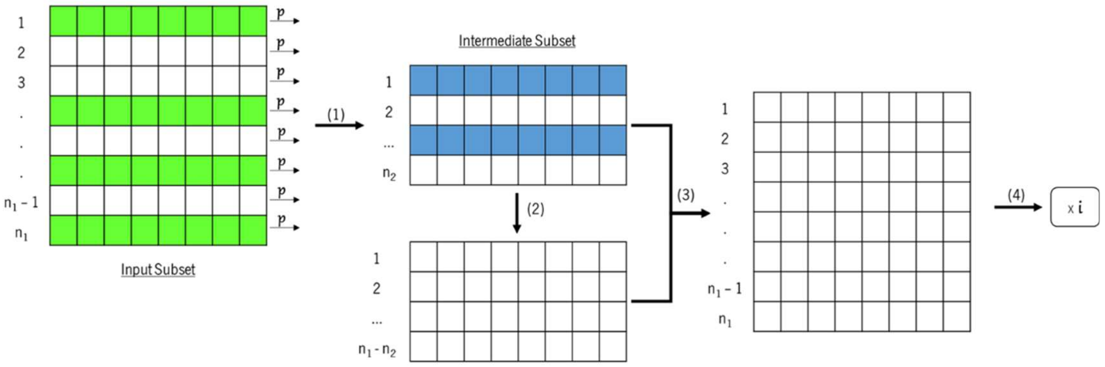


Figure 3 - Bootstrap Schematic. **(1)** Each entry in the input subset of size **n_1** has a probability (p) of being part of the intermediate subset of size **n_2**. In green are represented the entries that will form that subset. **(2)** **n_1 - n_2** entries (in blue) from the intermediate subset are randomly chosen to complete the intermediate subset, **(3)** creating a new subset of size **n_1**. **(4)** The process is iterated i times.

[**Figure 3 – (1)**], which is then completed by a replacement strategy; entries that are already part of the new subset are randomly chosen [**Figure 3 – (2,3)**]. This process is iterated i times. The bootstrap approach was applied with a p of 50% and 75% and an i of 1000. The arithmetic mean of the scores in every iteration was computed for each protein.

The pipeline followed is represented in **Figure 4**. PubMed was accessed to perform keyword-based queries and retrieve articles of interest (**Literature Search**). These were then analysed to extract data into excel spreadsheets and organize it according to the sample major characteristic and type of AF (**Information Extraction**). After this step, a scoring function was developed and implemented in *Python 3.5* and applied to the multiple subsets created (**Scoring Approach**). The final step was to determine the uncertainty of the developed method through bootstrapping (**Robustness Analysis**).

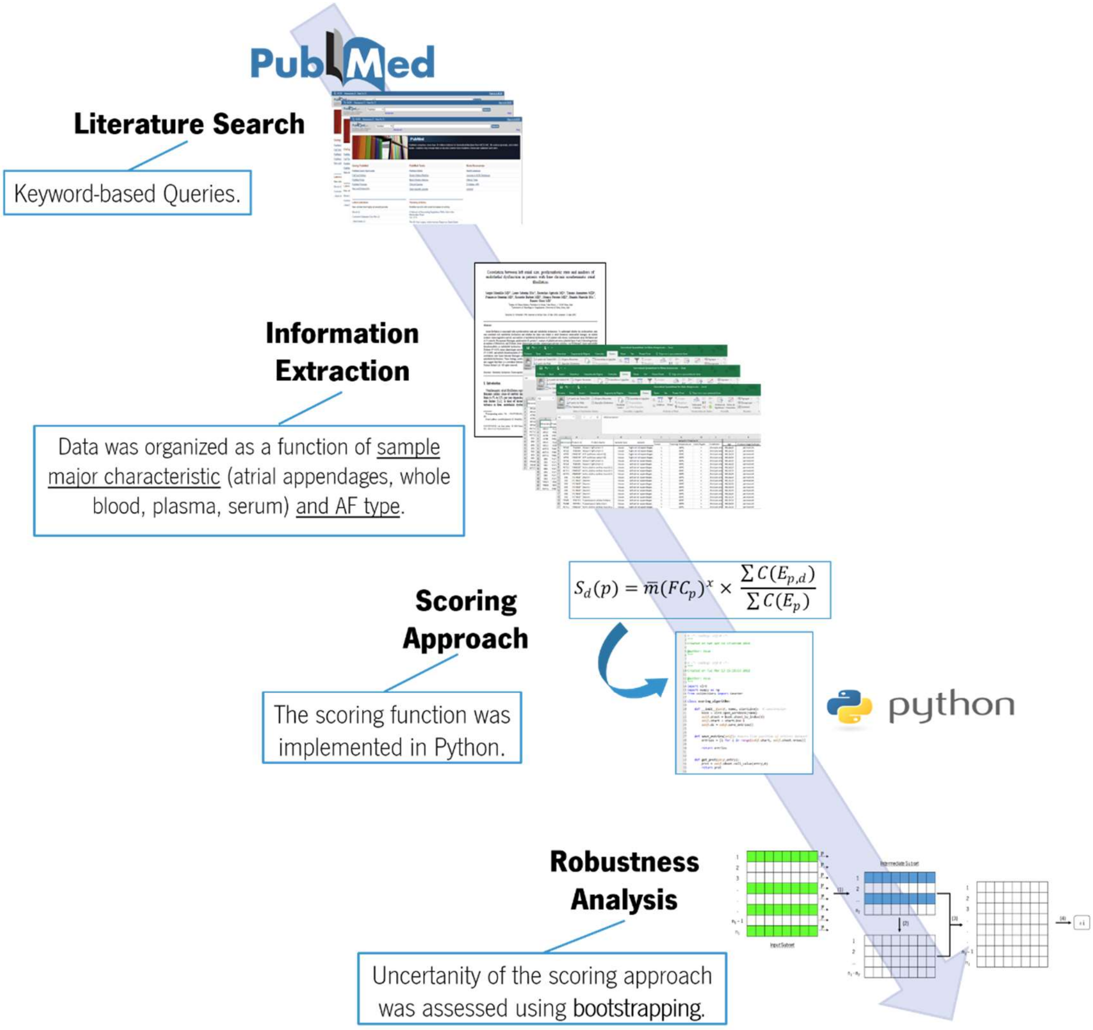

PubMed
Literature Search
Keyword-based Queries.
Information Extraction
Data was organized as a function of sample major characteristic (atrial appendages, whole blood, plasma, serum) and AF type.
Scoring Approach
$S_d(p) = \overline{m}(FC_p)^x \times \frac{\sum C(E_{p,d})}{\sum C(E_p)}$
python
The scoring function was implemented in Python.
Robustness Analysis
Uncertainty of the scoring approach was assessed using bootstrapping.
Input Subset
Intermediate Subset

Figure 4 - Pipeline of the methods used in the present work. PubMed was accessed to conduct keyword-based queries and retrieve relevant articles (**Literature Search**). Each article was analysed to extract relevant information into excel spreadsheets; data was organized according to the sample major characteristic and atrial fibrillation (AF) type (**Information Extraction**). A scoring function was developed, implemented in Python and applied to the created datasets (**Scoring Approach**). The uncertainty of the scoring approach was assessed by a bootstrap-based method (**Robustness Analysis**).

4. Results

The original dataset was composed of 712 entries. For each sample, the largest subsets were the "All" subsets, since they also include entries from the paroxysmal AF, persistent AF and permanent AF subsets of the corresponding sample, except for the atrial appendages subsets. The majority of the remaining subsets, however, had very few entries. The number of entries in each subset is represented in **Table 4**.

Table 4 – Distribution of the number of entries in each subset.

AF Types / Samples	Atrial Appendages	Whole Blood	Plasma	Serum
"All"	51	28	151	138
Paroxysmal AF	13	10	34	49
Persistent AF	56	7	32	64
Permanent AF	66	11	28	3
Postoperative new-onset AF	5	3	61	48
Postoperative AF recurrence	5	6	50	20

4.1 Scoring Systems

4.1.1 "All" Subsets – First-diagnosed, Paroxysmal, Persistent, Long-standing Persistent and Permanent AF

Regarding the atrial appendages-"All" dataset, 11/19 proteins were scored in the upwards direction and 8/19 proteins were scored in the opposite direction. Out of the 11 proteins scored in the direction "up", 10 had a score higher than one, the established threshold. All scores computed in the direction "down" passed the threshold. The proteins' UNIPROT code, full name, abbreviation, gene and respective scores, direction and number of entries in the subset are represented in **Supplemental Table 1**, ordered by the highest to the lowest scoring result. The fold-change range of the 10 highly-scored proteins in the direction "up" and of the eight proteins scored in the direction "down" is represented in **Figure 5 - (1)**. Every protein ranked in the upwards direction had a minimum fold-change value higher than one and every protein ranked in the downwards direction had a maximum fold-change value lower than one.

As for the whole blood-"All" dataset, four out of nine proteins/peptides had a computed score with "up" as the direction and only one score was computed with "down" as the direction. The latter reached the threshold, as did three scores calculated in the direction "up". The proteins whose score reached the threshold are present in **Supplemental Table 2** and their fold-change range in **Figure 5 – (2)**. All proteins whose score was computed in the direction “up” had a minimum fold-change value higher than one.

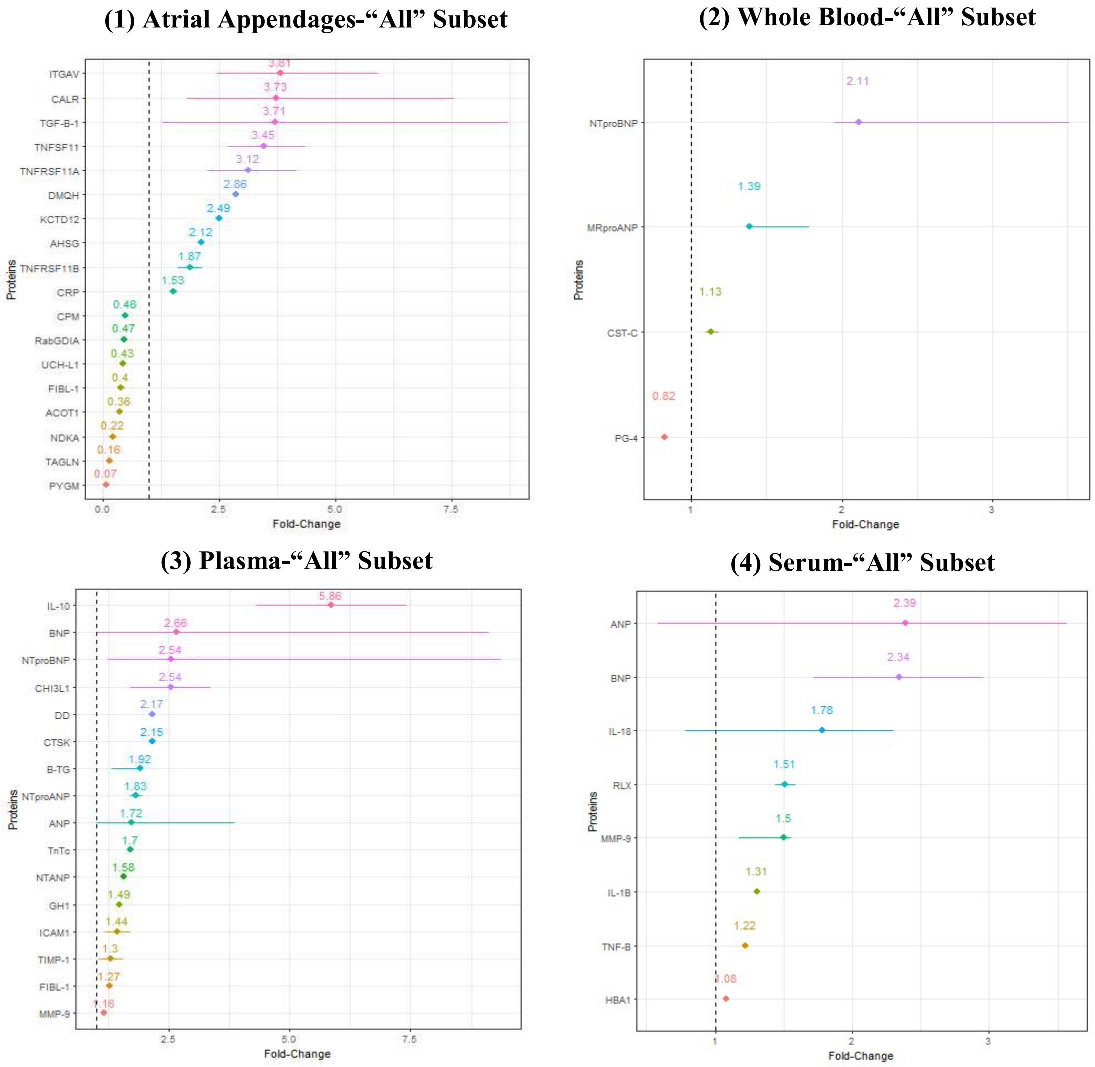


Figure 5 - Fold-change range of the proteins scored higher than one for each subset including all types of atrial fibrillation (AF). The median fold-change of each protein is represented. **1)** Atrial appendages-“All” subset; **2)** Whole blood-“All” subset; **3)** Plasma-“All” subset; **4)** Serum-“All” subset. **Abbreviations:** ACOT1, Acyl-coenzyme A thioesterase 1; AHSG, Alpha-2-HS-glycoprotein; ANP, Atrial Natriuretic Peptide; B-TG, Beta-thromboglobulin; BNP, Brain Natriuretic Peptide; CALR, Calreticulin; CHI3L1, Chitinase-3-like protein 1; CPM, Carboxypeptidase M; CRP, C-reactive protein; CST-C, Cystatin-C; CTSK, Cathepsin K; DD, D-dimer; DMQH, 5-demethoxyubiquinone hydroxylase, mitochondrial; FIBL-1, Fibulin-1; GH1, Glutathione hydrolase 1 proenzyme; HBA1, Hemoglobin subunit alpha; ICAM1, Intercellular adhesion molecule 1; IL-10, Interleukin-10; IL-18, Interleukin-18; IL-1B, Interleukin-1 beta; ITGAV, Integrin alpha-V; KCTD12, BTB/POZ domain-containing protein KCTD12; MMP-9, Matrix metalloproteinase-9; MRproANP, Mid-region pro-Atrial Natriuretic Peptide; NDKA, Nucleoside diphosphate kinase A; NTANP, N-terminal Atrial Natriuretic Peptide; NTproANP, N-terminal pro-Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; PG-4, Platelet glycoprotein 4; PYGM, Glycogen phosphorylase, muscle form; RLX, Relaxin; RabGDIA, Rab GDP dissociation inhibitor alpha; TAGLN, Transgelin; TGF-B-1, Transforming growth factor beta-1; TIMP-1, Metalloproteinase inhibitor 1; TNF-B, Lymphotoxin-alpha; TNFRSF11A, Tumor necrosis factor receptor superfamily member 11A; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFSF11, Tumor necrosis factor ligand superfamily member 11; TnTc, Troponin T, cardiac muscle; UCH-L1, Ubiquitin carboxyl-terminal hydrolase isozyme L1.

]

In the plasma-“All” subset, 24/37 and 1/37 scores were computed in the upwards and downwards direction, respectively. Of the 24 scores calculated in the upwards direction, 16 surpassed the established threshold (**Supplemental Table 3**); the only score computed in the direction “down” did not pass the threshold. Every protein/peptide ranked with $d = "up"$ had a fold-change value/values higher than one [**Figure 5 – (3)**].

When considering “serum” as the sample, 17 scores out of 23 were obtained with $d = "up"$ and three scores out of 23 with $d = "down"$. Only scores computed in the upwards direction passed the threshold, specifically eight of the scores (**Supplemental Table 4**). The minimum fold-change values for every protein/peptide, except for Atrial Natriuretic peptide (ANP), were higher than one [**Figure 5 – (4)**]

The selected top five biomarkers for this disease condition are represented in **Table 5**.

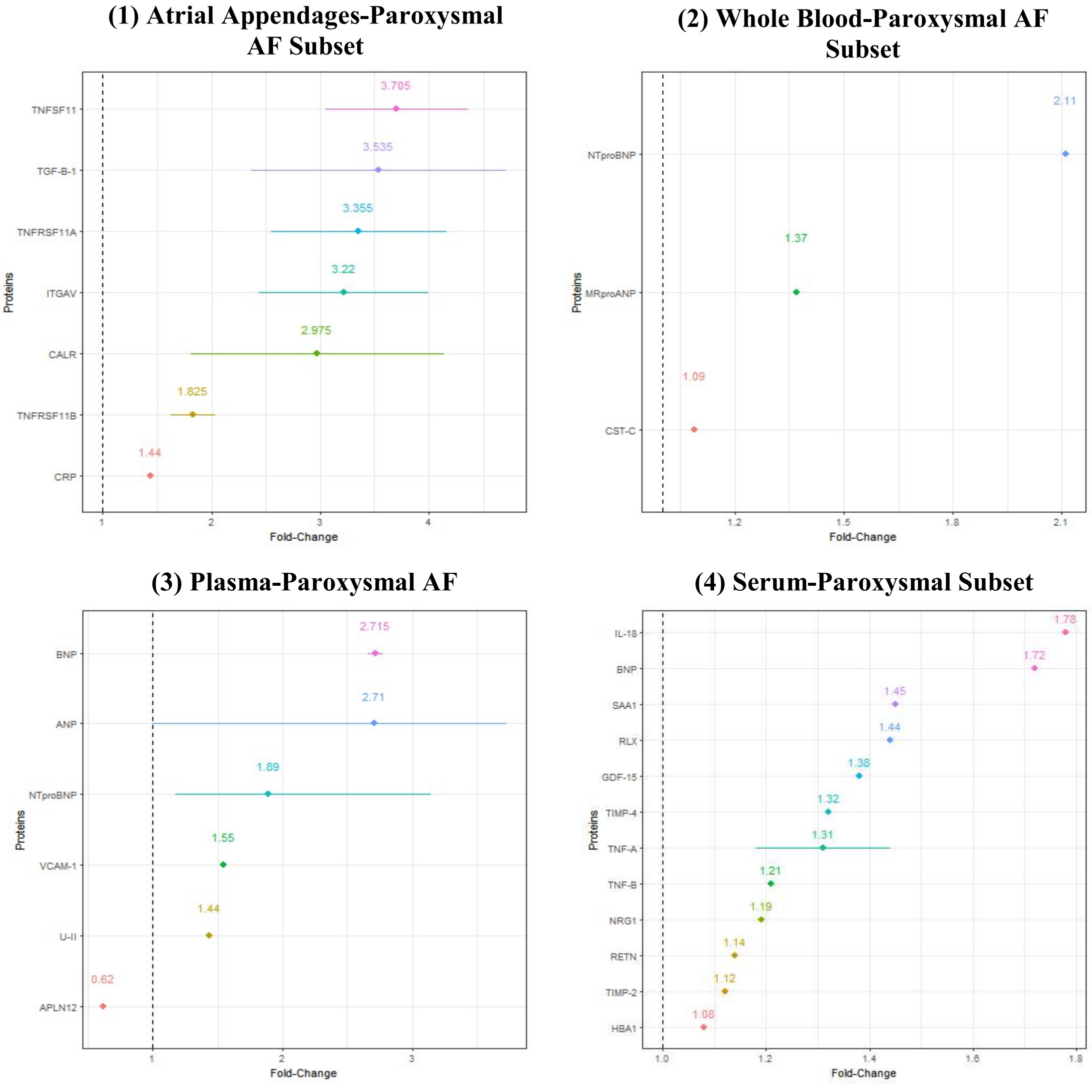
Table 5 – Top 5 proteins with the highest biomarker potential for Atrial Fibrillation.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries	Sample
P22301	Interleukin-10	IL-10	IL10	5.86	Up	2	Plasma
P36222	Chitinase-3-like protein 1	CHI3L1	CHI3L1	2.54	Up	2	Plasma
-	Brain Natriuretic Peptide	BNP	NPPB	2.34	Up	2	Serum
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	2.33	Up	10	Plasma
-	Brain Natriuretic Peptide	BNP	NPPB	2.22	Up	17	Plasma

4.1.2 Paroxysmal AF Subsets

Scores were only computed in the upwards direction for the atrial appendages-paroxysmal AF subset, more specifically seven out of seven and all passed the threshold (**Supplemental Table 5**), and whole blood-paroxysmal AF subset, namely four in four, out of which three passed the threshold [NTproBNP – 2.11, Mid-regional pro-Atrial Natriuretic peptide (MRproANP) – 1.37 and Cystatin-C (CST-C) - 1.09]. Every protein/peptide from both subsets had fold-change values higher than one [**Figure 6 – (1,2)**].

When considering plasma samples and paroxysmal AF as the condition, eight out of 13 proteins/peptides were ranked in the direction “up”, with six scores passing the threshold, and one protein was ranked in the opposite, with the corresponding score passing the threshold (**Supplemental Table 6**). The fold-change range is represented in **Figure 6 – (3)**. All fold-change values were higher than one for the proteins/peptides


(1) Atrial Appendages-Paroxysmal AF Subset

(2) Whole Blood-Paroxysmal AF Subset

(3) Plasma-Paroxysmal AF

(4) Serum-Paroxysmal Subset

Figure 6 - Fold-change range of the proteins scored higher than one for each subset with paroxysmal atrial fibrillation (AF) as the disease condition. The median fold-change of each protein is represented. **1)** Atrial appendages-paroxysmal AF subset; **2)** Whole blood-paroxysmal AF subset; **3)** Plasma-paroxysmal AF subset; **4)** Serum-paroxysmal AF subset. **Abbreviations:** ANP, Atrial Natriuretic Peptide; BNP, Brain Natriuretic Peptide; CALR, Calreticulin; CRP, C-reactive protein; CST-C, Cystatin-C; GDF-15, Growth/differentiation factor 15; HBA1, Hemoglobin subunit alpha; IL-18, Interleukin-18; ITGAV, Integrin alpha-V; MRproANP, Mid-region pro-Atrial Natriuretic Peptide; NRG1, neuregulin-1; NTproANP, N-terminal pro-Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; RETN, Resistin; RLX, Relaxin; SAA1, Serum amyloid A-1 protein; TGF-B-1, Transforming growth factor beta-1; TIMP-2, Metalloproteinase inhibitor 2; TIMP-4, Metalloproteinase inhibitor 4; TNF-A, Tumor necrosis factor; TNF-B, Lymphotoxin-alpha; TNFRSF11A, Tumor necrosis factor receptor superfamily member 11A; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFSF11, Tumor necrosis factor ligand superfamily member 11; U-II, Urotensin-2; VCAM-1, Vascular cell adhesion protein 1.

scored in the direction “up” and the fold-change of the protein scored in the direction “down” was lower than one.

In the serum-paroxysmal AF dataset, 30/42 and 2/42 scores were calculated in the “up” and “down” directions, respectively. With $d = "up"$ six scores surpassed the threshold and with $d = "down"$ no score passed the minimum value (**Supplemental Table 7**). All proteins/peptides had fold-change values higher than one [**Figure 6 – (4)**].

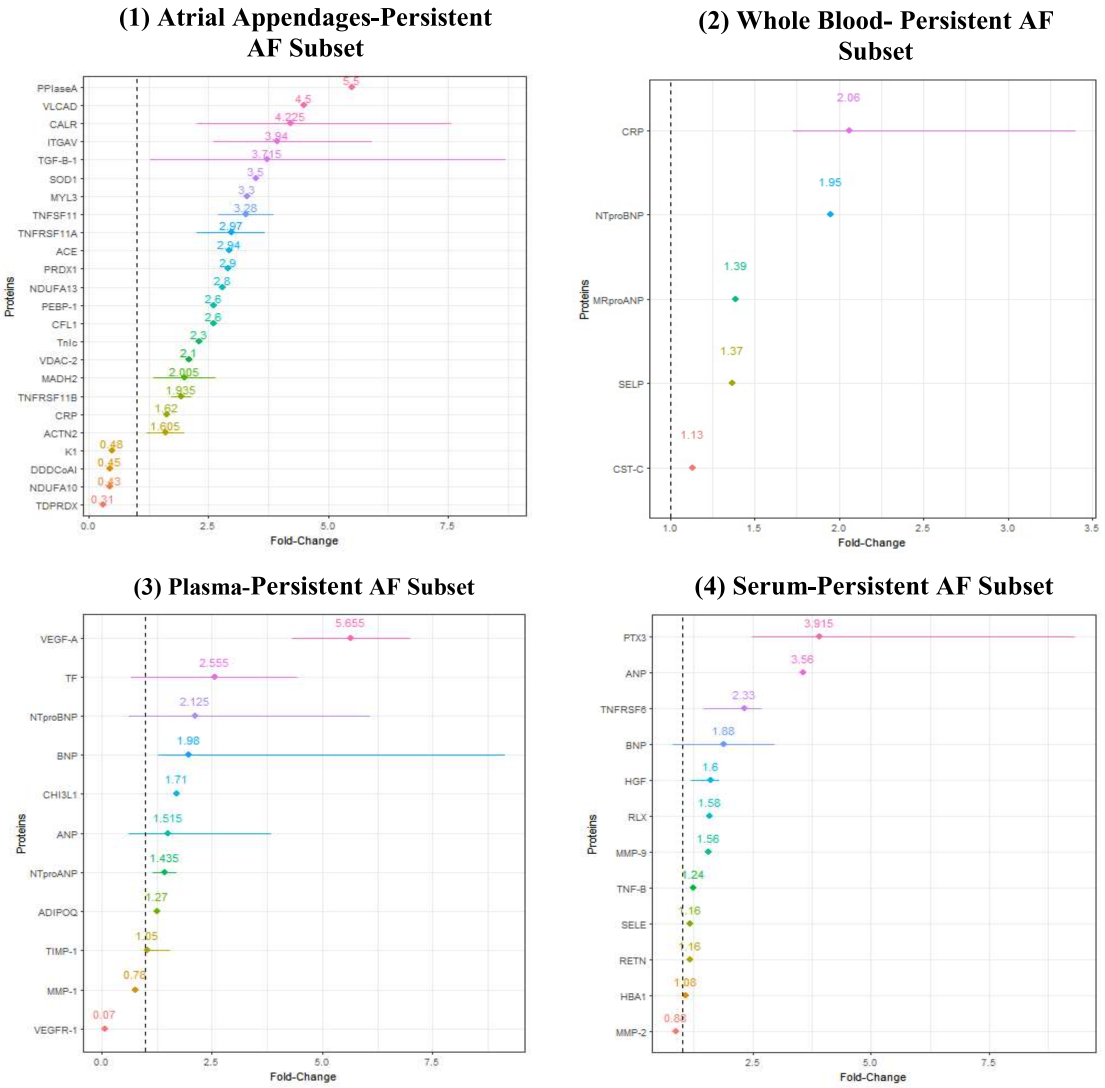
4.1.3 Persistent AF Subsets

Regarding the atrial appendages-persistent AF subset, 23/30 proteins were ranked when $d = "up"$, with 20 scores passing the threshold, and four proteins were ranked when $d = "down"$, with all scores passing the threshold (**Supplemental Table 8**). The proteins scored in the direction “down” only had one entry in the subset, which was defined as underexpressed (fold-change lower than one), and the proteins scored in the opposite direction only had fold-change values higher than one [**Figure 7 – (1)**].

In the whole blood-persistent AF subset, scores were only computed in the direction “up” and all five are higher than one (**Supplemental Table 9**). Every protein/peptide, except for CRP, was only present once in the subset but all fold-changes were higher than one [**Figure 7 – (2)**].

As for the plasma-persistent AF subset, 9/16 scores and 2/16 scores were computed in the “up” and “down” directions, respectively, and every score passed the threshold (**Supplemental Table 10**). The proteins/peptides which were scored with $d = "down"$ only had one entry each in the subset and a corresponding fold-change value lower than one. However, one protein and two peptides, namely Tissue Factor (TF), NTproBNP and ANP, which were scored in the direction “up”, had fold-changes ranging from values lower than one to values higher than one [**Figure 7 – (3)**].

In the serum-persistent AF subset, out of 26 only one score was calculated in the direction “down” and 16/26 were calculated in the opposite direction. The first passed the threshold and so did six of the scores computed in the direction “up” (**Supplemental Table 11**). The proteins/peptides which were ranked in the upwards direction had minimum fold-change values higher than one, except BNP, whose minimum fold-change value was lower than one [**Figure 7 – (4)**].



(1) Atrial Appendages-Persistent AF Subset

(2) Whole Blood- Persistent AF Subset

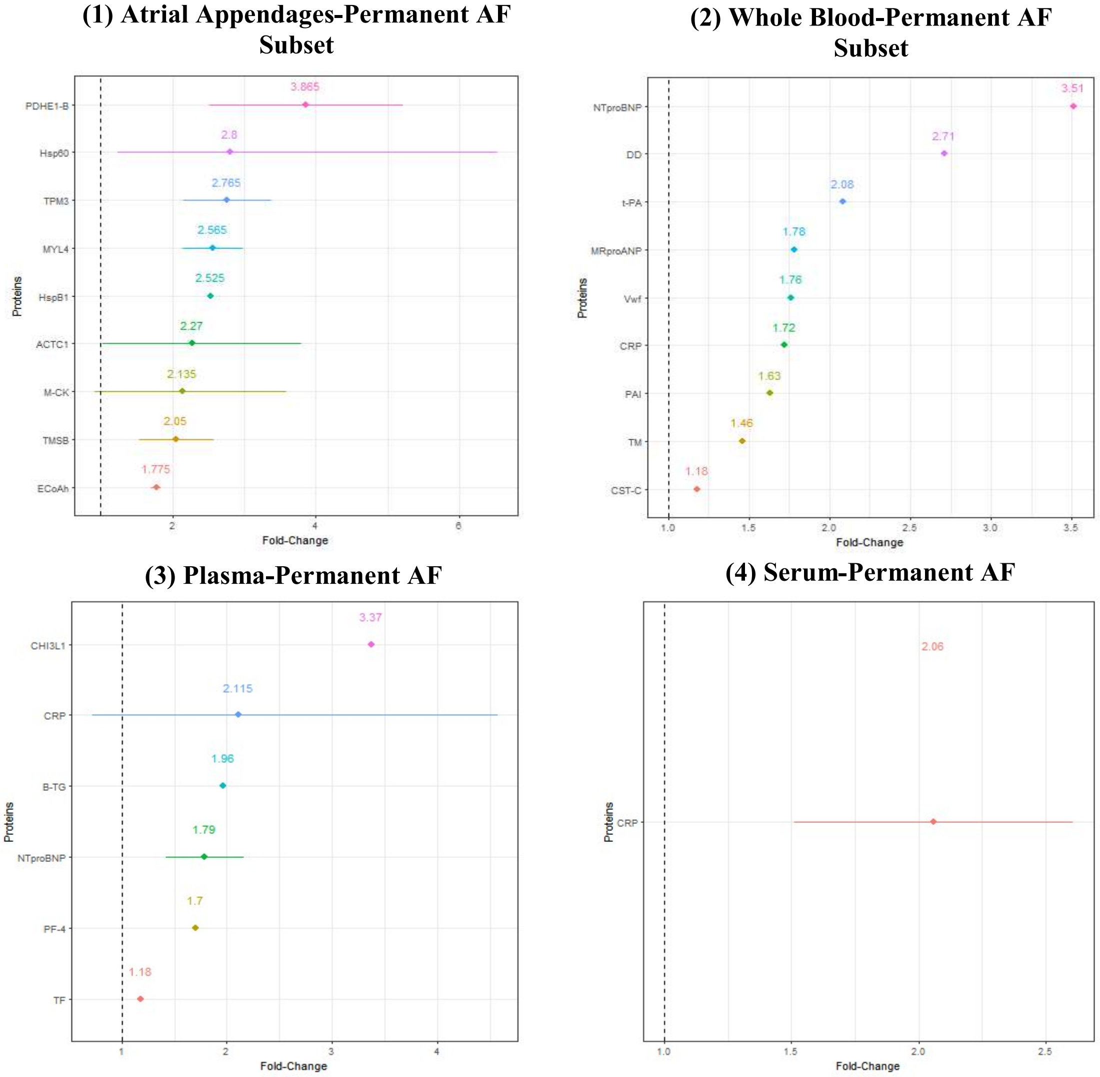
(3) Plasma-Persistent AF Subset

(4) Serum-Persistent AF Subset

Figure 7 - Fold-change range of the proteins scored higher than one for each subset with persistent atrial fibrillation (AF) as the disease condition. The median fold-change of each protein is represented. **1)** Atrial appendages-persistent AF subset; **2)** Whole blood-persistent AF subset; **3)** Plasma-persistent AF subset; **4)** Serum-persistent AF subset. **Abbreviations:** ACE, Angiotensin-converting enzyme; ACTN2, Alpha-actinin-2; ADIPOQ, Adiponectin; ANP, Atrial Natriuretic Peptide; BNP, Brain Natriuretic Peptide; CALR, Calreticulin; CFL1, Cofilin-1; CHI3L1, Chitinase-3-like protein 1; CRP, C-reactive protein; CST-C, Cystatin-C; DDDCoAI, Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase, mitochondrial; HBA1, Hemoglobin subunit alpha; HGF, Hepatocyte growth factor; ITGAV, Integrin alpha-V; K1, Keratin, type II cytoskeletal 1; MADH2, Mothers against decapentaplegic homolog 2; MMP-1, Interstitial collagenase; MMP-2, 72 kDa type IV collagenase; MMP-9, Matrix metalloproteinase-9; MRproANP, Mid-region pro-Atrial Natriuretic Peptide; MYL3, Myosin light chain 3; NDUFA10, NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 10, mitochondrial; NDUFA13, NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 13; NTproANP, N-terminal pro-Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; PEBP-1, Phosphatidylethanolamine-binding protein 1; PPIaseA, Peptidyl-prolyl cis-trans isomerase A; PRDX1, Peroxiredoxin-1; PTX3, Pentraxin-related protein PTX3; RETN, Resistin; RLX, Relaxin; SELE, E-selectin; SELP, P-selectin; SOD1, Superoxide dismutase Cu-Zn; TDPRDX, Thioredoxin-dependent peroxide reductase, mitochondrial; TF, Tissue factor; TGF-B-1, Transforming growth factor beta-1; TIMP-1, Metalloproteinase inhibitor 1; TNF-B, Lymphotoxin-alpha; TNFRSF11A, Tumor necrosis factor receptor superfamily member 11A; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFRSF6, Tumor necrosis factor receptor superfamily member 6; TNFSF11, Tumor necrosis factor ligand superfamily member 11; TnIc, Troponin I, cardiac muscle; VDAC-2, Voltage-dependent anion-selective channel protein 2; VEGF-A, Vascular endothelial growth factor A; VEGFR-1, Vascular endothelial growth factor receptor 1; VLCAD, Very long-chain specific acyl-CoA dehydrogenase, mitochondrial.

4.1.4 Permanent AF Subsets

In the atrial appendages dataset, scores were computed for every protein belonging to the subset (15) in the direction "up", with nine passing the threshold (**Supplemental Table 12**). All entries in the subset regarding these proteins had corresponding fold-change values higher than one, except one entry regarding M-CK [**Figure 8 – (1)**]. In the opposite direction, only Desmin (DES) was ranked, but the score did not pass the threshold.



(1) Atrial Appendages-Permanent AF Subset

(2) Whole Blood-Permanent AF Subset

(3) Plasma-Permanent AF

(4) Serum-Permanent AF

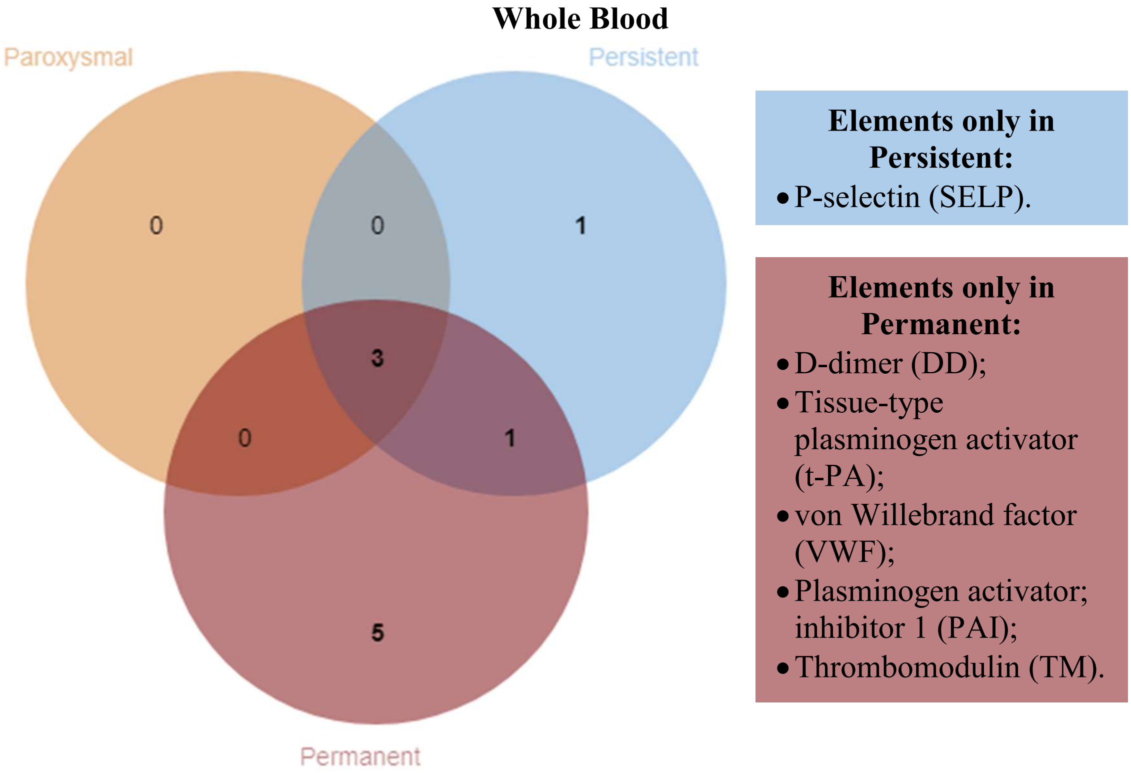
Figure 8 - Fold-change range of the proteins scored higher than one for each subset with permanent atrial fibrillation (AF) as the disease condition. The median fold-change of each protein is represented. **1)** Atrial appendages-permanent AF subset; **2)** Whole blood-permanent AF subset; **3)** Plasma-permanent AF subset; **4)** Serum-permanent AF subset. **Abbreviations:** ACTC1, Actin, alpha cardiac muscle 1; B-TG, Beta-thromboglobulin; CHI3L1, Chitinase-3-like protein 1; CRP, C-reactive protein; CST-C, Cystatin-C; DD, D-dimer; ECoAh, Enoyl-CoA hydratase, mitochondrial; Hsp60, 60 kDa heat shock protein, mitochondrial; M-CK, Creatine kinase M-type; MRproANP, Mid-region pro-Atrial Natriuretic Peptide; MYL4, Myosin light chain 4; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; PAI, Plasminogen activator inhibitor 1; PDHE1-B, Pyruvate dehydrogenase E1 component subunit beta, mitochondrial; PF-4, Platelet factor 4; TF, Tissue factor; TM, Thrombomodulin; TMSB, Tropomyosin beta chain; TPM3, Tropomyosin alpha-3 chain; Vwf, von Willebrand factor; t-PA, Tissue-type plasminogen activator.

In the remaining three subsets, no score was computed in the downwards direction. Nine out of 11 scores were calculated in the blood-permanent AF subset and all passed the threshold (**Supplemental Table 13**), eight in ten in the plasma-permanent AF subset, with six passing the threshold (**Supplemental Table 14**), and one out of two in the serum-permanent subset, which passed the threshold (CRP - 1.03). In the whole blood subset, CRP had a minimum fold-change value lower than one [**Figure 8 – (2)**], while in the plasma and serum subsets all fold-change values were higher than one [**Figure 8 – (3,4)**].

4.1.5 Top Five Biomarkers for Paroxysmal, Persistent and Permanent AF

Figures 9-11 show the proteins, which were highly-scored by the simple scoring approach, in comon between the paroxysmal, persistent and permanent AF types and concerning whole blood, plasma and serum as the major sample characteristic, respectively.

The top five selected biomarkers for the conditions paroxysmal, persistent and permanent AF are represented in **Tables 6-8**, respectively. No protein belonging to one of the top five was highly-scored in any of the other two conditions.



Elements only in Persistent:

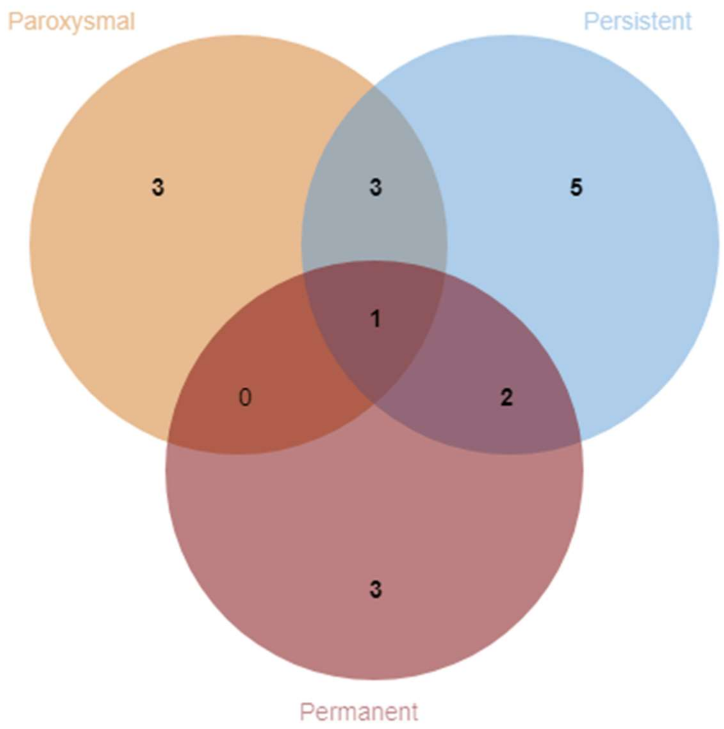
- P-selectin (SELP).

Elements only in Permanent:

- D-dimer (DD);
- Tissue-type plasminogen activator (t-PA);
- von Willebrand factor (VWF);
- Plasminogen activator; inhibitor 1 (PAI);
- Thrombomodulin (TM).

Figure 9 – Potential biomarkers only belonging to paroxysmal, persistent or permanent AF and whole blood as the sample.

Plasma



Elements only in Paroxysmal AF:

- Apelin-12 (APLN12);
- Vascular cell adhesion protein 1 (VCAM-1);
- Urotensin 2 (U-II).

Elements only in Persistent AF:

- Vascular Endothelial Growth Factor Receptor 1 (VEGFR-1);
- Vascular Endothelial Growth Factor A (VEGF-A);
- Metalloproteinase inhibitor 1 (TIMP-1);
- Interstitial collagenase (MMP-1);
- Adiponectin (ADIPOQ).

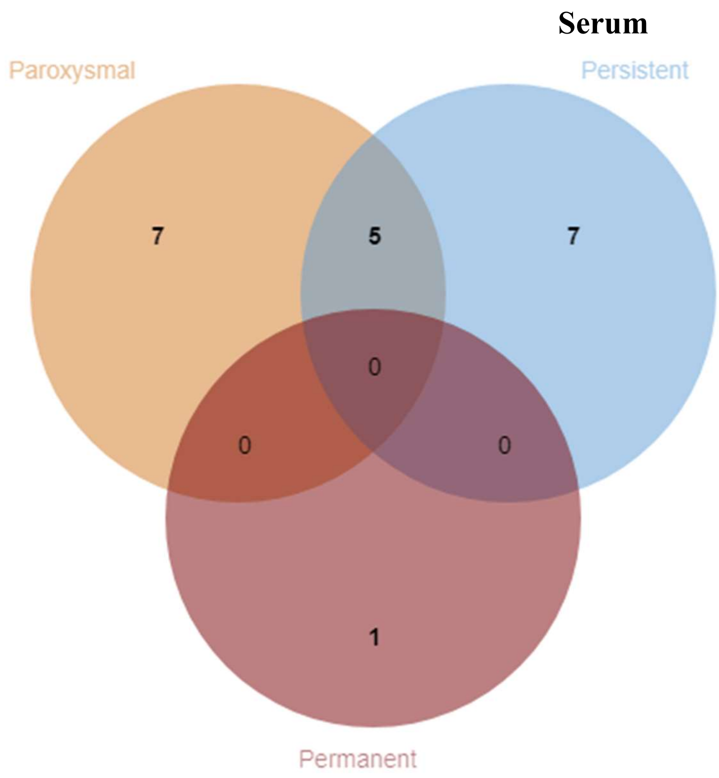
Elements only in Permanent AF:

- Beta-thromboglobulin (B-TG);
- Platelet factor 4 (PF-4);
- C-reactive protein (CRP).

Figure 10 – Potential biomarkers only belonging to paroxysmal, persistent or permanent AF and plasma as the sample.

Table 6 – Top 5 proteins with the highest biomarker potential for Paroxysmal Atrial Fibrillation.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries	Sample
Q14116	Interleukin-18	IL-18	IL18	1.78	Up	1	Serum
-	Apelin-12	APLN12	-	1.63	Down	1	Plasma
P19320	Vascular cell adhesion protein 1	VCAM-1	VCAM1	1.55	Up	1	Plasma
P0DJI8	Serum amyloid A-1 protein	SAA1	SAA1	1.45	Up	1	Serum
O95399	Urotensin-2	U-II	UTS2	1.44	Up	1	Plasma



Elements only in Paroxysmal AF:

- Interleukin-18 (IL-18);
- Serum amyloid A-1 protein (SAA1);
- Growth/differentiation factor 15 (GDF-15);
- Metalloproteinase inhibitor 4 (TIMP-4);
- Tumor necrosis factor (TNF-A);
- Neuroregulin-1 (NRG1)
- Tissue inhibitor metalloproteinase-2 (TIMP-2).

Elements only in Persistent AF:

- Atrial Natriuretic peptide (ANP):
- Pentraxin-related protein PTX3 (PTX3);
- Hepatocyte growth factor (HGF);
- Matrix metalloproteinase-9 (MMP-9);
- Tumor necrosis factor receptor superfamily member 6 (TNFRSF6);
- E-selectin (SELE);
- 72 kDa type IV collagenase (MMP-2).

Elements only in Permanent AF:

- C-reactive protein (CRP).

Figure 11 – Potential biomarkers only belonging to paroxysmal, persistent or permanent AF and serum as the sample.

Table 7 – Top 5 proteins with the highest biomarker potential for Persistent Atrial Fibrillation.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries	Sample
P17948	Vascular endothelial growth factor receptor 1	VEGFR-1	FLT1	7.16	Down	2	Plasma
-	Atrial Natriuretic Peptide	ANP	NPPA	3.56	Up	1	Serum
P15692	Vascular endothelial growth factor A	VEGF-A	VEGFA	2.83	Up	2	Plasma
P26022	Pentraxin-related protein PTX3	PTX3	PTX3	1.96	Up	4	Serum
P14210	Hepatocyte growth factor	HGF	HGF	1.60	Up	3	Serum

Table 8 – Top 5 proteins with the highest biomarker potential for Permanent Atrial Fibrillation.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries	Sample
-	D-dimer	DD	-	2.71	Up	1	Blood
P00750	Tissue-type plasminogen activator	t-PA	PLAT	2.08	Up	1	Blood
-	Beta-thromboglobulin	B-TG	PPBP	1.96	Up	1	Plasma
P04275	von Willebrand factor	VWF	VWF	1.76	Up	1	Blood
P02776	Platelet factor 4	PF-4	PF4	1.70	Up	1	Plasma

4.1.6 Postoperative new-onset AF Subsets

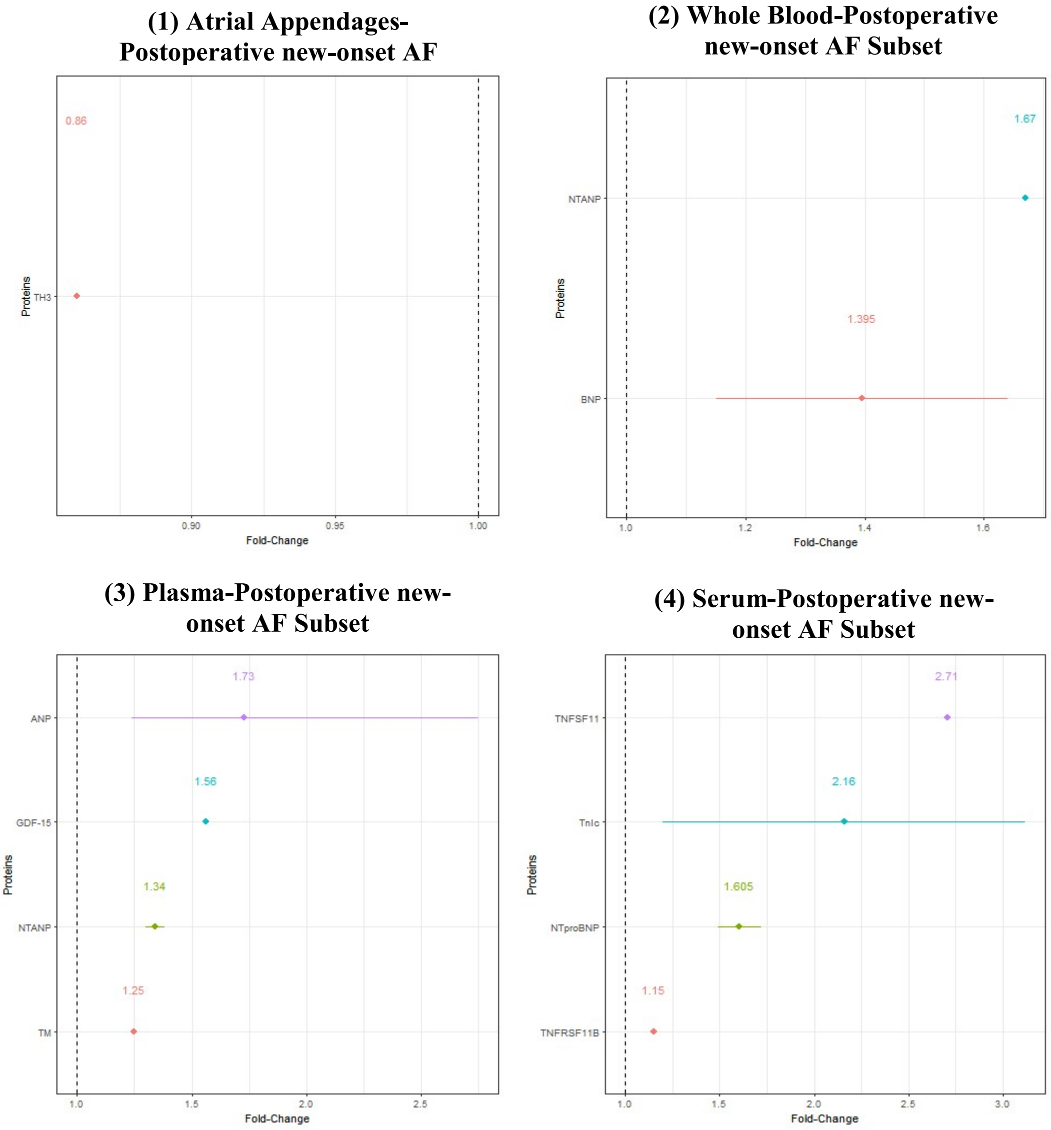
Only Triiodothyronine (TH3) had a score computed (1.17) in the atrial appendages-postoperative new-onset AF subset of a total of five proteins, specifically in the direction “down”, and the score passed the threshold.

The whole blood and the plasma-postoperative new-onset AF subsets only had scores computed when $d = "up"$. In the blood subset both peptides were ranked, namely N-terminal Atrial Natriuretic peptide (NTANP) and BNP, with 1.67 and 1.40 as the respective scores. Nine proteins/peptides out of 15 were ranked in the plasma subset, with four passing the threshold (**Supplemental Table 15**).

Regarding the serum-postoperative new-onset AF dataset, five in 32 scores (16 in each direction) were calculated; one in the downwards direction, which did not pass the threshold and four in the upwards direction, which passed the threshold (**Supplemental Table 16**).

Figure 12 shows that every fold-change value for every protein/peptide, whose score passed the threshold, in the four subsets was higher than one.

The selected top five biomarkers for this disease condition are represented in **Table 9**.



(1) Atrial Appendages-Postoperative new-onset AF

(2) Whole Blood-Postoperative new-onset AF Subset

(3) Plasma-Postoperative new-onset AF Subset

(4) Serum-Postoperative new-onset AF Subset

Figure 12 - Fold-change range of the proteins scored higher than one for each subset with postoperative new-onset atrial fibrillation (AF) as the disease condition. The median fold-change of each protein is represented. **1)** Atrial appendages-postoperative new-onset AF subset; **2)** Whole blood-postoperative new-onset AF subset; **3)** Plasma-postoperative new-onset AF subset; **4)** Serum-postoperative new-onset AF subset. **Abbreviations:** ANP, Atrial Natriuretic Peptide; BNP, Brain Natriuretic Peptide; GDF-15, Growth/differentiation factor 15; NTANP, N-terminal Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; TH3, Triiodothyronine; TM, Thrombomodulin; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFSF11, Tumor necrosis factor ligand superfamily member 11; TnIc, Troponin I, cardiac muscle.

Table 9 – Top 5 proteins with the highest biomarker potential for Postoperative New-onset Atrial Fibrillation.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries	Sample
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	2.71	Up	1	Serum
-	N-terminal Atrial Natriuretic peptide	NTANP	NPPA	1.67	Up	1	Blood
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	1.60	Up	2	Serum
Q99988	Growth/differentiation factor 15	GDF-15	GDF15	1.56	Up	1	Plasma
-	Atrial Natriuretic Peptide	ANP	NPPA	1.44	Up	6	Plasma

4.1.7 Postoperative AF Recurrence Subsets

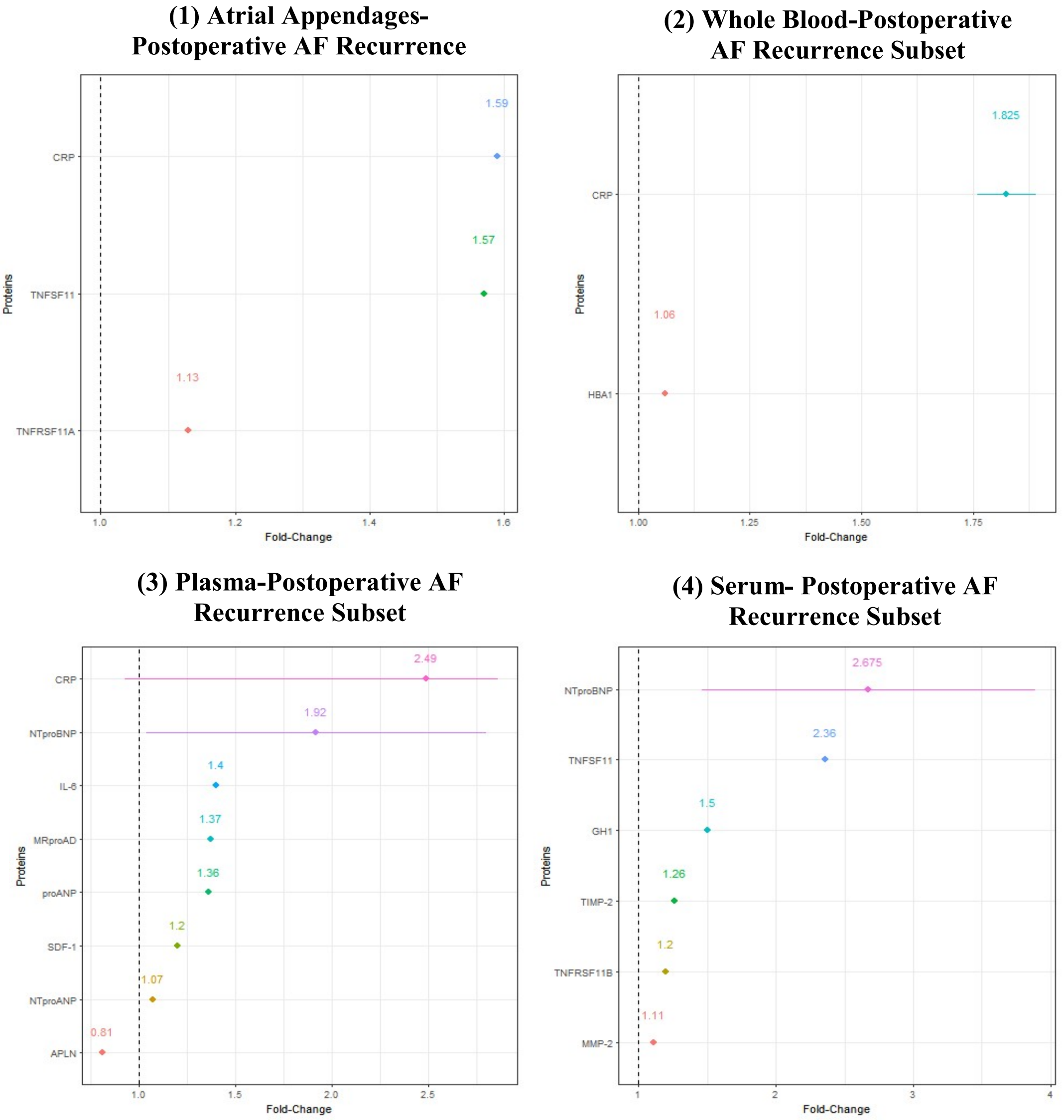
No score was computed in the direction “down” in the atrial appendages, whole blood and serum-postoperative AF recurrence subsets. Three of five scores were computed in the atrial appendages-postoperative AF recurrence subset and all three passed the defined minimum value [CRP - 1.59, Tumor necrosis factor ligand superfamily member 11 (TNFSF11) - 1.57 and Tumor necrosis factor receptor superfamily member 11A (TNFRSF11A) - 1.13]. The fold-change values of the corresponding proteins are all higher than one [**Figure 13 – (1)**].

In the whole blood-postoperative AF recurrence, two out of three proteins were ranked but only Hemoglobin subunit alpha (HBA1) ’s score, namely 1.06, passed the threshold. The fold-change value is represented in **Figure 13 – (2)**.

As for the plasma-postoperative AF recurrence, 11/15 proteins were scored in the direction “up”, but only seven passed the threshold, and one was scored in the direction “down” and passed the threshold (**Supplemental Table 17**). Nonetheless, CRP, which was scored in the upwards direction, has a minimum fold-change value lower than one, while the remaining proteins/peptides scored in the same direction only present fold-change values higher than one [**Figure 13 – (3)**].

In the serum-postoperative AF recurrence dataset, seven out of 12 scores were calculated and six reached the threshold (**Supplemental Table 18**). The fold-change values of the corresponding proteins/peptides are all higher than one [**Figure 13 – (4)**].

The selected top five biomarkers for this disease condition are represented in **Table 10**.



(1) Atrial Appendages-Postoperative AF Recurrence

(2) Whole Blood-Postoperative AF Recurrence Subset

(3) Plasma-Postoperative AF Recurrence Subset

(4) Serum- Postoperative AF Recurrence Subset

Figure 13 - Fold-change range of the proteins scored higher than one for each subset with postoperative atrial fibrillation (AF) recurrence as the disease condition. The median fold-change of each protein is represented. **1)** Atrial appendages-postoperative AF recurrence AF subset; **2)** Whole blood-postoperative AF recurrence AF subset; **3)** Plasma-postoperative AF recurrence AF subset; **4)** Serum-postoperative AF recurrence AF subset. **Abbreviations:** APLN, Apelin; CRP, C-reactive protein; GH1, Glutathione hydrolase 1 proenzyme; HBA1, Hemoglobin subunit alpha; IL-6, Interleukin-6; MMP-2, 72 kDa type IV collagenase; MRproAD, mid-regional pro-adrenomedullin; NTproANP, N-terminal pro-Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; SDF-1, Stromal cell-derived factor 1; TIMP-2, Metalloproteinase inhibitor 2; TNFRSF11A, Tumor necrosis factor receptor superfamily member 11A; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFSF11, Tumor necrosis factor ligand superfamily member 11; proANP, pro-Atrial Natriuretic Peptide.

4.1.8 Biological Functions

The biological functions associated with the highly-scored proteins/peptides were retrieved; proteins/peptides were divided into 10 groups based on their major biological function (**Table 11**): metabolism, regulation of ion molecules handling/concentration, atrial contraction and muscle fibres formation/organization, fibrosis, inflammation, fibrinolysis/fibrinogenesis and coagulation, vasoconstriction/vasodilation, oxidative stress, apoptosis and others.

Table 10 – Top 5 proteins with the highest biomarker potential for Postoperative Atrial Fibrillation Recurrence.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries	Sample
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	2.36	Up	1	Serum
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	1.92	Up	2	Plasma
P02741	C-reactive protein	CRP	CRP	1.66	Up	7	Plasma
P02741	C-reactive protein	CRP	CRP	1.59	Up	1	Atrial Appendages
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	1.57	Up	1	Atrial Appendages

Table 11 – Distribution of the highly-scored proteins according to their major biological function.

Protein	Biological Function
• Pyruvate dehydrogenase E1 component subunit beta, mitochondrial (PDHE1-B) • Enoyl CoA hydratase (ECoAh) • Adiponectin (ADIPOQ) • NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 13 (NDUFA13) • Very long-chain specific acyl-CoA dehydrogenase, mitochondrial (VLCAD) • NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 10, mitochondrial (NDUFA10) • Acyl-coenzyme A thioesterase 1 (ACOT1) • Glutathione hydrolase 1 proenzyme (GH1) • Creatine Kinase type M (M-CK) • Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase, mitochondrial (DDDCoAI) • Glycogen phosphorylase, muscle form (PYGM) • Nucleoside diphosphate kinase A (NDKA)	**Metabolism**
• Triiodothyronine (TH3) • Fibulin-1 (FIBL-1) • Calreticulin (CALR) • Integrin alpha-V (ITGAV) • Stromal cell-derived factor-1 (SDF-1) • Voltage-dependent anion-selective channel protein 2 (VDAC-2) • BTB/POZ domain-containing protein KCTD12 (KCTD12) • Rab GDP dissociation inhibitor alpha (RabGDIA)	**Regulation of Ion Molecules Handling/Concentration**

Table 11 – Distribution of the highly-scored proteins according to their major biological function (continued).

Protein	Biological Function
• Transgelin (TAGLN) • Tropomyosin alpha-3 chain (TPM3) • Tropomyosin beta chain (TMSB) • Myosin light chain 4 (MYL4) • Actin, alpha cardiac muscle 1 (ACTC1) • Apelin-12 (APLN12) • Apelin (APLN) • Troponin I, cardiac muscle (TnIc) • Troponin T, cardiac muscle (TnT • Alpha-actinin-2 (ACTN2) • Cofilin-1 (CFL-1) • Myosin light chain 3 (MYL3) • Ubiquitin carboxyl-terminal hydrolase isozyme L1 (UCH-L1) • Keratin, type II cytoskeletal 1 (K1)	**Atrial Contraction and Muscle Fibres Formation/Organization**
• Metalloproteinase inhibitor 2 (TIMP-2) • Metalloproteinase inhibitor 4 (TIMP-4) • Transforming growth factor beta-1 (TGF-B-1) • Growth/differentiation factor 15 (GDF-15) • Mothers against decapentaplegic homolog 2 (MADH2) • Brain Natriuretic peptide (BNP) • N-terminal Brain Natriuretic peptide (NTproBNP) • Pro-Brain Natriuretic peptide (proBNP) • Relaxin (RLX) • Tumor necrosis factor beta (TNF-B)	**Fibrosis**
• Peroxiredoxin-1 (PRDX1) • Superoxide dismutase Cu-Zn (SOD1) • Thioredoxin-dependent peroxide reductase, mitochondrial (TDPRDX) • 5-demethoxyubiquinone hydroxylase, mitochondrial (DMQH) • Ceruloplasmin (CP)	**Oxidative Stress**
• Cystatin-C (CST-C) • Tumor necrosis factor receptor superfamily 6 (TNFRSF6)	**Apoptosis**

Table 11 – Distribution of the highly-scored proteins according to their major biological function (continued).

Protein	Biological Function
• Resistin (RETN) • Serum amyloid A-1 protein (SAA1) • Interleukin-1 beta (IL-1B) • Tumor necrosis factor receptor superfamily member 11A (TNFRSF11A) • Tumor necrosis factor receptor superfamily member 11B (TNFRSF11B) • Tumor necrosis factor ligand superfamily member 11 • C-reactive protein (CRP) • Chitinase-3-like protein 1 (CHI3L1) • Platelet factor 4 (PF-4 • P-selectin (SELP) • Pentraxin-related protein PTX3 (PTX3) • Alpha-2-HS-glycoprotein (AHSG) • Vascular endothelial growth factor A (VEGF-A) • Vascular endothelial growth factor receptor 1 (VEGFR-1) • Vascular cell adhesion protein 1 (VCAM-1) • Interleukin-18 (IL-18) • Tumor necrosis factor A (TNF-A) • Interleukin-10 (IL-10) • Interleukin-6 (IL-6) • Platelet glycoprotein 4 (PG-4) • Intercellular adhesion molecule 1 (ICAM-1)	**Inflammation**
• D-dimer (DD) • Plasminogen activator inhibitor 1 (PAI) • Beta-thromboglobulin (B-TG) • Von Willebrand factor (VWF) • Tissue factor (TF) • Thrombomodulin (TM)	**Fibrinolysis/Fibrinogenesis and Coagulation**
• Urotensin-2 (U-II) • Angiotensin-converting enzyme (ACE) • Mid-region pro-Adrenomedullin (MRproAD) • Atrial Natriuretic peptide (ANP) • N-terminal Atrial Natriuretic peptide (NTANP) • N-Terminal pro-Atrial Natriuretic peptide (NTproANP) • Mid-region pro-Atrial Natriuretic peptide (MRproANP)	**Vasoconstriction/vasodilation**
• 60 kDa heat shock protein, mitochondrial (Hsp60) • Heat shock protein beta-1 (HspB1) • Peptidyl-prolyl cis-trans isomerase A (PPIaseA) • Mucin-16 (MUC-16) • Hemoglobin subunit alpha (HBA1) • Neuregulin-1 (NRG1) • Carboxypeptidase M (CPM) • E-selectin (SELE) • Hepatocyte growth factor (HGF) • Phosphatidylethanolamine-binding protein 1 (PEBP-1)	**Others**

4.2 Bootstrap

Some subsets were not sufficiently large or had very few proteins with very few entries each for the bootstrap with a p of 50% to be successfully applied in either direction, namely the blood subsets with paroxysmal AF, persistent AF or permanent AF as the disease conditions, the atrial appendages and blood subsets with postoperative new-onset AF and postoperative AF recurrence as the conditions, the serum-permanent AF subset and the serum-postoperative new-onset AF subset. The bootstrap was applied with a p of 75% to the blood subsets with paroxysmal AF, persistent AF or permanent AF as the disease conditions and the results are presented further ahead. The remaining subsets had very few entries for a bootstrap method to be applied, regardless of the value of p. The atrial appendages and blood-postoperative AF recurrence subsets, the blood-postoperative new-onset AF and the serum-permanent AF subsets had proteins which were scored by the simple scoring approach in the direction “up” but not in the direction “down”. However, they only have 2-3 entries and, thus, the bootstrap is not applied in these cases. The atrial appendages-postoperative new-onset AF subset had one protein scored in the direction “down”, but this protein was only represented by one entry and, as such, the bootstrap was not applied.

4.2.1 Bootstrap with $p = 50\%$

4.2.1.1 “All” Subsets – First-diagnosed, Paroxysmal, Persistent, Long-standing Persistent and Permanent AF

Regardless of the sample, all subsets respecting to the condition “All” were sufficiently large for the bootstrap with a p of 50% to be applied in the direction “up”.

In the results of the bootstrap approach applied to the atrial appendages-“All” subset, seven proteins were successfully scored but only six had a mean score higher than one. The proteins with a mean score higher than the threshold were also highly-scored with the scoring approach without iterations (**Supplemental Table 19**). All proteins defined as underexpressed had very few entries in the subset, thus the bootstrap was not applied in the direction “down”. The range of scores for each protein is represented in **Figure 14 – (1)**. The minimum score for every protein is zero, which means that, for each protein, no entry or only “unchanged entries” were selected in at least one iteration.

As for the whole blood-“All” subset, only two peptides [NTproBNP – 2.20 and MRproANP – 1.31] had a mean score higher than one, specifically in the direction “up”. The same peptides were also highly-scored with the simple scoring approach. **Figure 14**

– **(2)** shows the score range of the two proteins; each score range started at zero, which indicates that, for each protein, no entry was selected in at least one entry for the final subset. There was only one protein scored with $d = "down"$ by the simple scoring approach and it was only represented by one entry in the subset. Likewise, the bootstrap was not applied in the direction “down”.

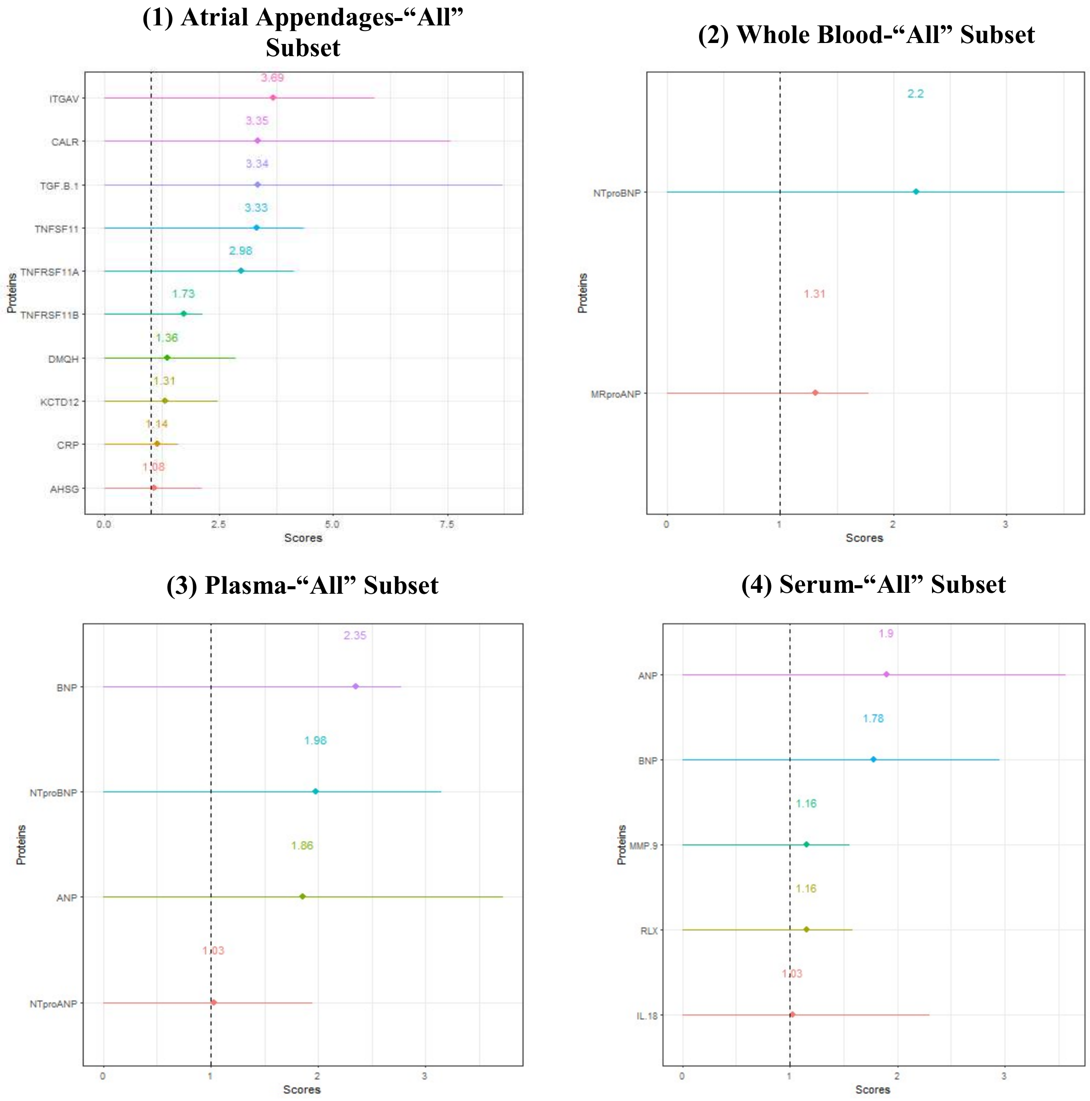


Figure 14 - Score range of the proteins with a mean score higher than one for the subsets with “All” as the disease condition. The mean score of each protein is represented. **1)** Atrial appendages-“All” AF subset; **2)** Whole blood-“All” subset; **3)** Plasma-“All” subset; **4)** Serum-“All” subset. **Abbreviations:** AHSG, Alpha-2-HS-glycoprotein; ANP, Atrial Natriuretic Peptide; B-TG, Beta-thromboglobulin; BNP, Brain Natriuretic Peptide; CALR, Calreticulin; CHI3L1, Chitinase-3-like protein 1; CRP, C-reactive protein; CTSK, Cathepsin K; DD, D-dimer; DMQH, 5-demethoxyubiquinone hydroxylase, mitochondrial; ICAM1, Intercellular adhesion molecule 1; IL-10, Interleukin-10; IL-18, Interleukin-18; ITGAV, Integrin alpha-V; KCTD12, BTB/POZ domain-containing protein KCTD12; MMP-9, Matrix metalloproteinase-9; MRproANP, Mid-region pro-Atrial Natriuretic Peptide; NTproANP, N-terminal pro-Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; RLX, Relaxin; TF, Tissue factor; TGF-B-1, Transforming growth factor beta-1; TNFRSF11A, Tumor necrosis factor receptor superfamily member 11A; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFSF11, Tumor necrosis factor ligand superfamily member 11.

In the plasma-“All” subset, 11 scores passed the threshold with $d = "up"$ and the respective proteins were also ranked as potential biomarkers in the previous section, except for TF (**Supplemental Table 20**). The score range is represented in **Figure 14 – (3)**. The minimum score for every protein was zero, which means that, for each protein, no entry, only “unchanged entries” or only entries defined as “N/A” were selected in at least one iteration. There was only one protein, which only had one entry in the subset, whose score was calculated with the scoring approach without iterations and with $d = "down"$. Thus, the bootstrap was not applied in the direction “down”.

In the serum-“All” subset, five scores passed the threshold (**Supplemental Table 21**), specifically in the direction “up”, and the respective proteins also had scores calculated with the simple scoring approach higher than one. The range of scores of the bootstrap method is represented in **Figure 14 – (4)**. The minimum score for each protein was zero, which indicates that, for each protein, no entries, only entries defined as unchanged or “N/A” or, in the case of ANP, only the entry defined as “underexpressed” were selected for the final subset. No score passed the threshold with $d = "down"$.

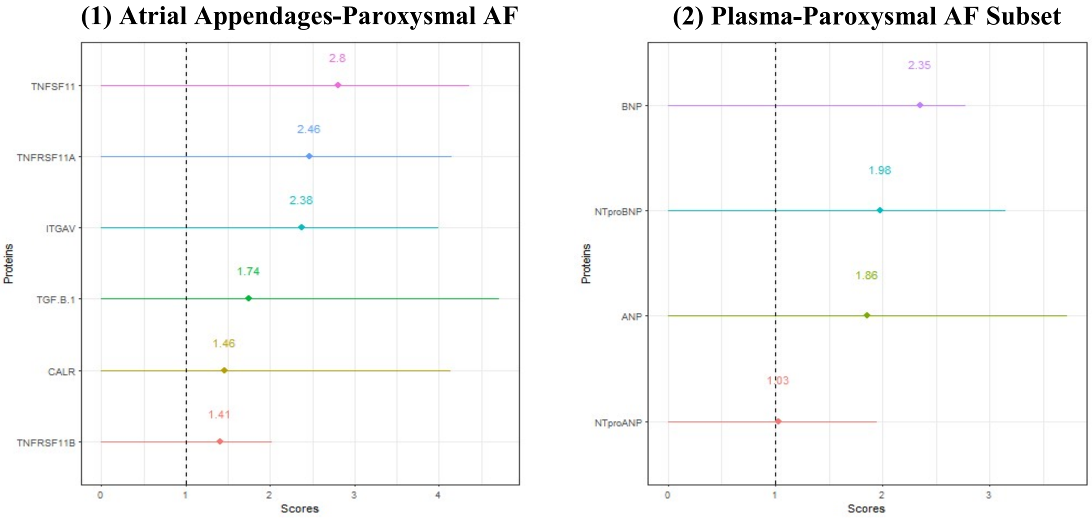
4.2.1.2 Paroxysmal AF Subsets

In the atrial appendages-paroxysmal AF subset, six proteins had a computed score higher than one in the direction “up” (**Supplemental Table 22**). The same proteins were also highly-scored by the simple approach and only CRP’s score did not pass the threshold after the bootstrap. No protein was scored with $d = "down"$ by the simple scoring method, so the bootstrap method was not applied in this direction.

In the plasma-paroxysmal AF subset, four scores were higher than one (**Supplemental Table 23**) and two did not pass the threshold. The proteins, whose score passed the threshold in the bootstrap system, were also considered potential biomarkers according to the simple scoring approach. There was only one protein scored by the simple scoring algorithm in the direction “down” and it was only represented by one entry; thus, the bootstrap was not applied in the direction “down”.

In both subsets, each score range started at zero [**Figure 15 – (1,2)**], which indicates that, for each protein, no entry or only unchanged entries were selected in at least one of the iterations.

As for the serum-paroxysmal AF subset’s results, no mean score passed the threshold in the direction “up”. Additionally, there were only two proteins which were scored by the simple scoring method in the direction “down”, but scores did not pass the threshold. Hence, the bootstrap was not applied in this direction.


(1) Atrial Appendages-Paroxysmal AF

(2) Plasma-Paroxysmal AF Subset

Figure 15 – Score range of the proteins with a mean score higher than one for the subsets with paroxysmal atrial fibrillation (AF) as the disease condition. The mean score of each protein is represented. **1)** Atrial appendages-paroxysmal AF subset; **2)** Plasma-paroxysmal AF subset. **Abbreviations:** ANP, Atrial Natriuretic Peptide; BNP, Brain Natriuretic Peptide; CALR, Calreticulin; ITGAV, Integrin alpha-V; NTproANP, N-terminal pro-Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; TGF-B-1, Transforming growth factor beta-1; TNFRSF11A, Tumor necrosis factor receptor superfamily member 11A; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFSF11, Tumor necrosis factor ligand superfamily member 11.

4.2.1.3 Persistent AF Subsets

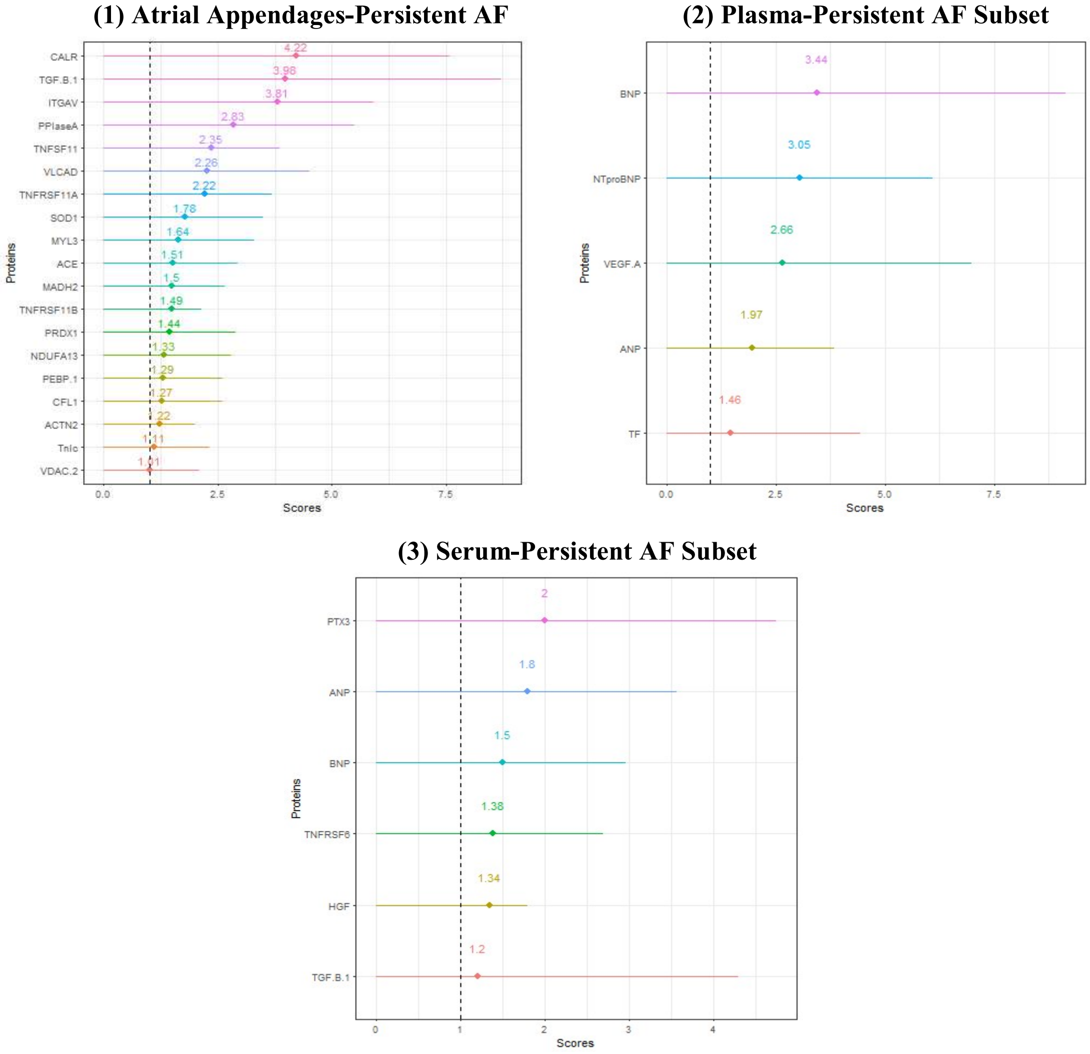
Regardless of the sample, there were very few proteins scored in the direction “down”, thus the bootstrap was not applied with $d = "down"$

In the atrial appendages-persistent AF dataset, 18 scores passed the threshold (**Supplemental Table 24**) and the corresponding proteins also had scores higher than one in the previous section. The minimum score computed for each protein was zero [**Figure 16 – (1)**], which means that, in at least one iteration, no entry was selected for each protein.

In the plasma-persistent AF subset, five proteins had an average score higher than one (**Supplemental Table 25**) and the same proteins also had high scores derived from the simple scoring calculation. Additionally, the minimum score found in the iterations for every protein was zero [**Figure 16 – (2)**], indicating that no entry or only entries defined as “N/A” were selected for each protein in at least one iteration.

As for the serum-persistent AF subset, six proteins had mean scores that passed the threshold (**Supplemental Table 26**). Transforming growth factor beta-1 (TGF-B-1) was the only protein which did not have a score, computed with the scoring system without iterations, higher than one. The range of scores in the multiple iterations is shown in

Figure 16 – (3). All ranges started at zero, indicating that in at least one iteration no entry or only entries defined as "N/A" were selected for the final subset, for each protein.


(1) Atrial Appendages-Persistent AF

(2) Plasma-Persistent AF Subset

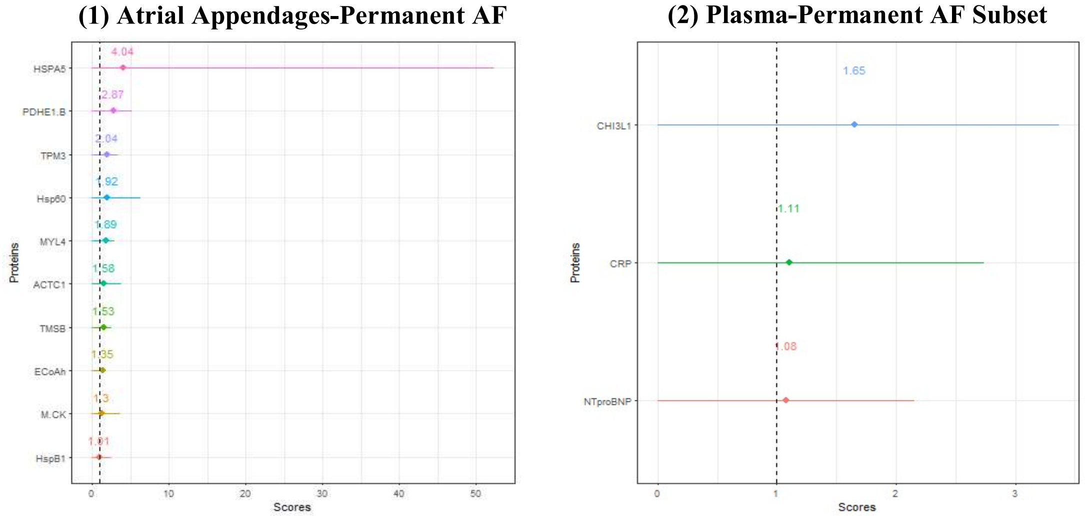
(3) Serum-Persistent AF Subset

Figure 16 – Score range of the proteins with a mean score higher than one for the subsets with persistent atrial fibrillation (AF) as the disease condition. The mean score of each protein is represented. **1)** Atrial appendages-persistent AF subset; **2)** Plasma-persistent AF subset; **3)** Serum-persistent subset. **Abbreviations:** ACE, Angiotensin-converting enzyme; ACTN2, Alpha-actinin-2; ANP, Atrial Natriuretic Peptide; BNP, Brain Natriuretic Peptide; CALR, Calreticulin; CFL1, Cofilin-1; HGF, Hepatocyte growth factor; ITGAV, Integrin alpha-V; MADH2, Mothers against decapentaplegic homolog 2; MYL3, Myosin light chain 3; NDUFA13, NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 13; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; PEBP-1, Phosphatidylethanolamine-binding protein 1; PPIaseA, Peptidyl-prolyl cis-trans isomerase A; PRDX1, Peroxiredoxin-1; PTX3, Pentraxin-related protein PTX3; SOD1, Superoxide dismutase Cu-Zn; TF, Tissue factor; TGF-B-1, Transforming growth factor beta-1; TNFRSF11A, Tumor necrosis factor receptor superfamily member 11A; TNFRSF11B, Tumor necrosis factor receptor superfamily member 11B; TNFRSF6, Tumor necrosis factor receptor superfamily member 6; TNFSF11, Tumor necrosis factor ligand superfamily member 11; TnIc, Troponin I, cardiac muscle; VEGF-A, Vascular endothelial growth factor A; VLCAD, Very long-chain specific acyl-CoA dehydrogenase, mitochondrial.

4.2.1.4 Permanent AF Subsets

In the atrial appendages-permanent AF subset, ten proteins had a mean score higher than one (**Supplemental Table 27**) in the direction "up" and the same proteins, except for Endoplasmic reticulum chaperone BiP (HSPA5), also had high scores calculated with the simple scoring formula. The minimum score for every protein was zero [**Figure 17 – (1)**], which means that, for each protein, no entry or only "unchanged entries" were selected in at least one iteration. Moreover, the range for HSPA5 varied between zero and 50 due to the entry with a corresponding fold-change of 52. With $d = "down"$, the mean score of DES, the only protein belonging to the subset in this condition, did not pass the threshold.

In the plasma-permanent AF dataset, only three proteins had scores which passed the threshold [Chitinase-3-like protein 1 (CHI3L1) – 1.65, CRP – 1.11 and NTproBNP – 1.08]. The same three proteins were also considered potential biomarkers in the previous section. The range of scores is represented in **Figure 17 – (2)** and had a minimum of zero, meaning that no entry or only entries defined as unchanged or "N/A" were selected in at least one iteration for each protein. No protein was scored with $d = "down"$, so the bootstrap method was not applied in this direction.



(1) Atrial Appendages-Permanent AF

(2) Plasma-Permanent AF Subset

Figure 17 – Score range of the proteins with a mean score higher than one for the subsets with permanent atrial fibrillation (AF) as the disease condition. The mean score of each protein is represented. **1)** Atrial appendages-permanent AF subset; **2)** Plasma-permanent AF subset. **Abbreviations:** ACTC1, Actin, alpha cardiac muscle 1; CHI3L1, Chitinase-3-like protein 1; CRP, C-reactive protein; ECoAh, Enoyl-CoA hydratase, mitochondrial; HSPA5, Endoplasmic reticulum chaperone BiP; Hsp60, 60 kDa heat shock protein, mitochondrial; M-CK, Creatine kinase M-type; MYL4, Myosin light chain 4; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; PDHE1-B, Pyruvate dehydrogenase E1 component subunit beta, mitochondrial; TMSB, Tropomyosin beta chain; TPM3, Tropomyosin alpha-3 chain.

4.2.1.5 Postoperative new-onset AF Subsets

Regarding the postoperative new-onset AF condition, the bootstrap system was only successfully applied with a p of 50% to the plasma subset with $d = "up"$. Three scores passed the established threshold [ANP – 1.49, BNP – 1.22 and NTANP – 1.04], out of which two of the corresponding proteins (ANP and NTANP) also had high scores computed with the simple scoring method. All score ranges started at zero (**Figure 18**), which indicates that no entry or only “unchanged entries” were selected in at least one iteration for each protein. No proteins were scored with $d = "down"$ by the simple scoring algorithm. Hence, the bootstrap was not applied in this direction.

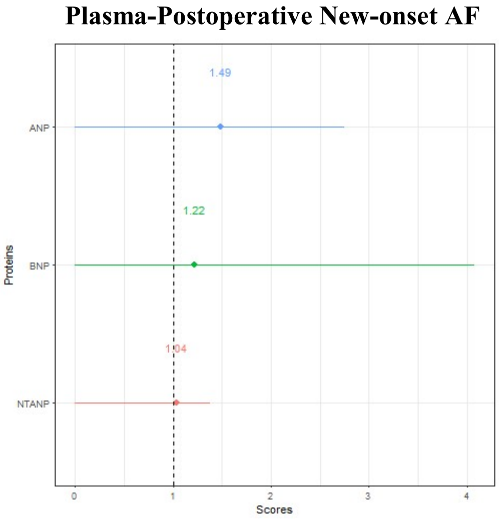


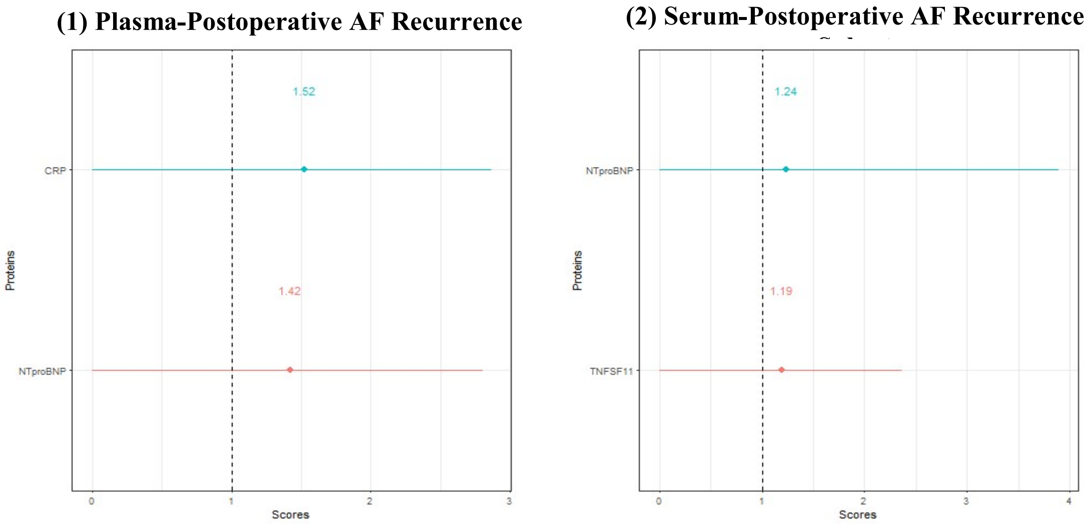
Figure 18 – Score range of the proteins with a mean score higher than one for the plasma-postoperative new-onset atrial fibrillation (AF) subset. **Abbreviations:** ANP, Atrial Natriuretic Peptide; BNP, Brain Natriuretic Peptide; NTANP, N-terminal Atrial Natriuretic Peptide.

4.2.1.6 Postoperative AF Recurrence Subsets

In the plasma-postoperative AF recurrence subset, only two proteins had scores that surpassed the threshold (CRP – 1.52 and NTproBNP – 1.42) and they also presented scores higher than one calculated with the approach without iterations. After eliminating the entries corresponding to proteins which were not scored by the simple scoring approach with $d = "down"$, the input subset for the bootstrap in this direction only had one entry, thus the bootstrap was not applied.

In the serum-postoperative AF recurrence subset, NTproBNP and TNFSF11 had mean scores higher than one, 1.24 and 1.19, respectively. These proteins also had high scores calculated with the simple scoring method. No proteins were scored with $d = "down"$ by the simple scoring algorithm, hence the bootstrap was not applied in this direction.

For both subsets, the minimum score for each protein was zero [**Figure 19 – (1,2)**], which indicates that no entry or only entries defined as unchanged or “N/A” were selected in at least one iteration.


(1) Plasma-Postoperative AF Recurrence

(2) Serum-Postoperative AF Recurrence

Figure 19 – Score range of the proteins with a mean score higher than one for the subsets with postoperative atrial fibrillation (AF) recurrence as the disease condition. The mean score of each protein is represented. **1)** Plasma-postoperative AF recurrence subset; **2)** Serum-postoperative AF recurrence subset. **Abbreviations:** CRP, C-reactive protein; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; TNFSF11, Tumor necrosis factor ligand superfamily member 11.

4.2.2 Bootstrap 75%

As previously mentioned, the bootstrap approach with a p of 75% was applied to some subsets, more specifically, the blood subsets with paroxysmal AF, persistent AF or permanent AF as the disease conditions and the serum-postoperative new-onset AF subset. Regardless of the subset, no protein was scored by the simple scoring system in the direction “down” and, thus, the bootstrap was not applied in this direction. As such, the following results were obtained with $d = "up"$.

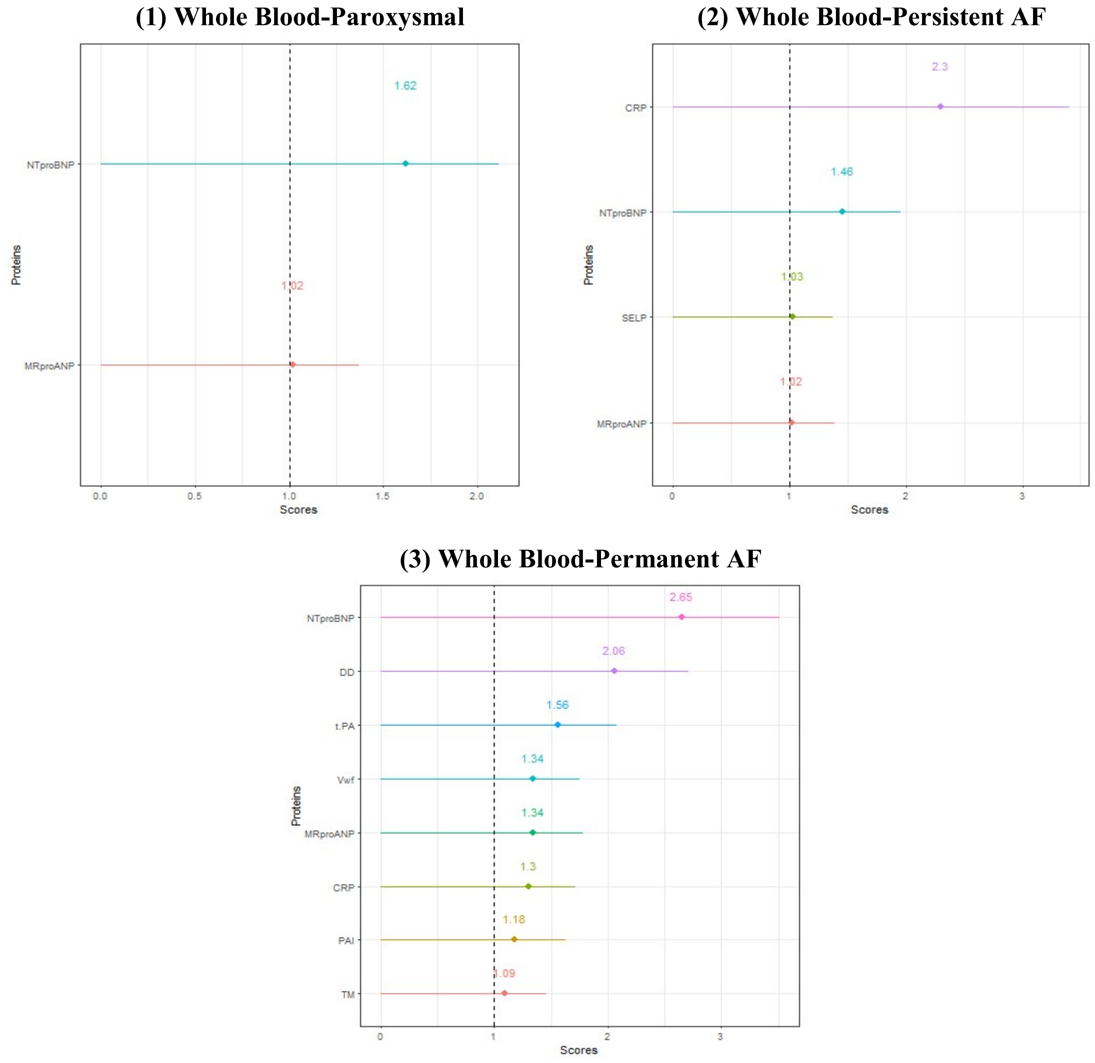
Regarding the whole blood-paroxysmal AF subset, NTproBNP and MRproANP had a mean score higher than one (1.62 and 1.02, respectively). These two proteins also had high scores computed with the scoring approach without iterations.

Only four scores passed the threshold in the whole blood-persistent AF dataset (**Supplemental Table 28**) and the corresponding proteins were also considered potential biomarkers according to the results obtained with the simple scoring method.

As for the whole blood-permanent AF subset, eight proteins presented scores higher than one (**Supplemental Table 29**). The same proteins also hade scores that passed the threshold in the previous section.

For the three subsets, each range of scores started at zero [**Figure 20 – (1,2,3)**], indicating that no entry was selected in at least one iteration for each protein.

Finally, three scores passed the threshold in the serum-postoperative new-onset AF [TNFSF11 – 2.09, NTproBNP - 1.51 and Troponin I, cardiac muscle (TnIc) – 1.28] and the corresponding proteins were also highly-scored by the method without iterations. The minimum score for each protein was zero (**Figure 21**), indicating that no entry or, in the case of TnIc, only "unchanged entries" were selected for the final subset in at least one iteration for each protein.



(1) Whole Blood-Paroxysmal

(2) Whole Blood-Persistent AF

(3) Whole Blood-Permanent AF

Figure 20 – Score range of the proteins with a mean score higher than one for the subsets with whole blood as the sample and paroxysmal, persistent and permanent atrial fibrillation (AF) as the disease condition. The mean score of each protein is represented. **1)** Whole blood-paroxysmal AF subset; **2)** Whole blood-persistent AF subset; **3)** Whole blood-permanent subset. **Abbreviations:** CRP, C-reactive protein; DD, D-dimer; MRproANP, Mid-region pro-Atrial Natriuretic Peptide; NTproBNP, N-terminal pro-Brain Natriuretic Peptide; PAI, Plasminogen activator inhibitor 1; SELP, P-selectin; TM, Thrombomodulin; Vwf, von Willebrand factor; t-PA, Tissue-type plasminogen activator.

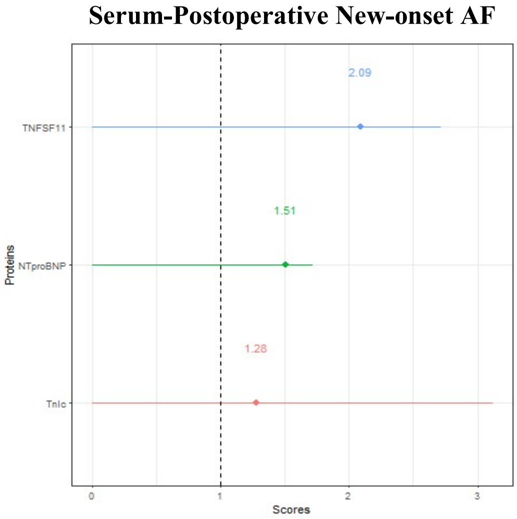


Figure 21 - Score range of the proteins with a mean score higher than one for the serum-postoperative new-onset atrial fibrillation (AF) subset. **Abbreviations:** NTproBNP, N-terminal pro-Brain Natriuretic Peptide; TNFSF11, Tumor necrosis factor ligand superfamily member 11; TnIc, Troponin I, cardiac muscle.

5. Discussion

AF's molecular profiling might improve diagnostic and prognostic disease management. Protein biomarkers are important beacons in this process, hence a gauging benchmark is required to determine their relative potential. We developed a bioinformatics-oriented scoring function aimed at weighing the importance of proteins/peptides and mitigating the limitations of the currently known scores.

5.1 Scoring Approach Performance and Validation

One of the main issues when choosing a potential biomarker for a given disease is the coherence between findings. For instance, CRP was measured in the serum of patients with paroxysmal AF in nine different articles, which in total contributed with 11 entries to the original dataset. CRP was classified as overexpressed in five entries, as unchanged in other five and as "N/A" in one entry. The simple scoring approach seems to poorly score proteins in this situation, which indicates that the scoring system correctly reflects the incoherence between studies. This is evidenced by the coherence between findings for each protein regardless of the subset, which can be observed in **Figures 5-13**. For the proteins defined and scored as overexpressed and as potential biomarkers for AF, if the minimum fold-change encountered is higher than one, then there is complete consistency between findings. For the proteins defined and scored as underexpressed and as potential biomarkers for AF, if the highest fold-change encountered is lower than one, then there is complete consistency between findings. However, there are still some cases in which proteins had scores higher than the threshold. Actin, alpha cardiac muscle 1 (ACTC1) was found to be either increased or unchanged in the atrial appendages of individuals with permanent AF. Although differences between groups in the "unchanged entries" did not reach statistical significance, the fold-change was higher than one, which might indicate that differences were reaching statistical significance. Moreover, the high fold-change pulls the score up, leading to scores higher than the threshold.

Additionally, the highly-scored proteins were divided into nine major groups based on their major biological function: metabolism, regulation of ion molecules concentration, atrial contraction and muscle fibres formation/organization, fibrosis, inflammation, fibrinogenesis/fibrinolysis and coagulation, vasoconstriction/vasodilation, oxidative stress and apoptosis. Changes in the expression levels of proteins involved in either of these biological functions/alterations seem to agree with AF's pathophysiology and clinical presentation. In fact, the overexpression or underexpression of proteins involved in these processes might contribute to AF's development and/or maintenance or represent a biological response to changes induced by the disease itself. The raised levels of proteins involved in metabolic processes most likely act as a response to the increase in energy demand during the early phases of AF, which was observed in experimental animals [134,135], and might continue through latter phases of AF and be caused by the augmented atria and the cardiac contractile dysfunction.

On the other hand, lowered levels of certain proteins might act in order to counterbalance the raised production of energy and of certain proteins. The overexpression or underexpression of proteins/peptides involved in the handling of ionic molecules, especially Ca^{2+}, K^+ and Na^+, might initiate and maintain AF, through the generation of DADs and EADs, and result in poor atrial contraction, also observed in AF. Additionally, the overexpression/underexpression of molecules involved in actin filament binding and sliding might change the contractile properties of the atria, contributing, once again, to contractile dysfunction. A variety of proteins participates in proteolysis/reorganization of the extracellular matrix, degradation of fibronectin and deposition of several collagen types, which might lead to the accumulation of fibrotic tissue, one of the major mechanisms responsible for cardiac remodelling [136] and characteristic of AF.

Regarding inflammation, growing evidence suggests that inflammation is associated with AF and that it plays a part in AF's pathophysiology [137,138]. Furthermore, AF itself can induce inflammation during the remodelling process, perpetuating the disease [139]. AF patients exhibit a high thrombogenic state, which raises the risks of stroke and thromboembolism [140]. It seems plausible for this elevated prothrombotic state to originate from the dysregulation of proteins which participate in the events concerning haemostasis. In response to the low ejection fraction of AF individuals due to accumulation of blood in the heart, which in turn results from poor contraction, raised levels of vasoconstrictor proteins might be produced to boost the amount of blood that reaches the heart and contradict the low ejection fraction. In opposition, in response to the overexpression of vasoconstrictors, vasodilator proteins might be released to counterbalance the effects produced by the first. Evidence supports that oxidative stress occurs in the hearts of subjects with AF and that it may play a role in remodelling [141–143]. This pathological condition might result from the upregulation or down-regulation of proteins with oxidant/anti-oxidant functions.

Finally, the apoptotic process was observed in AF and associated with the Ca^{2+} overload in the heart, which leads to rapid activation [178]. Nonetheless, the overexpression/underexpression of proteins regulating apoptosis might also be the cause of this event. As such, the results obtained by the simple scoring approach seem consistent with AF, which gives credibility to the method developed.

However, because the scoring approach does not take into account the number of times a protein has been studied in AF, a variety of proteins was highly-scored despite only having been studied once or few times in AF. Thus, more studies are needed to better understand if these proteins are actually changed in AF and to find their true scoring value.

To validate our scoring system in face of different realities, the method was applied in bootstrap form. Most of the highly-scored proteins by the bootstrap system were also highly-scored by the simple method, meaning that in light of different situations the developed method is effective. These proteins represent the ones with the most precise score. Furthermore, the bootstrap approach did not score as highly proteins which were only studied once or few times, with some scores not even passing the threshold. This happens since each iteration represents a different "reality" and proteins with more entries are most likely to be represented by at least one entry in the final subset in each iteration, compared to proteins with only one or few entries, which end up not being selected in multiple iterations. Still, some proteins which only appear once or few times in a certain subset continue to have high average scores for one or both of the following reasons: 1) they present high or low fold-change values, depending of the direction, which result in high scores in the iterations in which the entry was selected and end up compensating for the iterations in which the entry was not selected (e.g.: 5-demethoxyubiquinone hydroxylase, mitochondrial (DMQH) was highly-scored despite only having one entry in the atrial appendages-"All" subset because it had a fold-change of 2.86); 2) the subset is small and, so, there is a higher chance for the entry to be selected (e.g.: MRproANP had a fold-change of 1.37 in the whole blood-paroxysmal AF subset, which was only composed of 11 entries). Hence, these two cases confirm that the scoring system is limited by or depend on the amount of existing studies.

Notwithstanding, if perceived as an actual scoring method, there are four cases in which the bootstrap approach did not perform well and scored highly proteins, which were poorly scored by the simple method. The first case concerns HSPA5, which was measured six times in the atrial appendages of permanent AF individuals and defined as overexpressed in just two of the measurements. However, one of the “unchanged entries” had a corresponding fold-change of 2.47 and the “overexpressed entries” had fold-changes of 52.45 and 7.21. The high fold-changes, especially the first, originated high scores in the iterations where at least one of these entries was selected. Furthermore, the fold-change of 52.42 was very high and could be considered an outlier or result from measurement errors.

As for the second case, TF had four entries in the plasma-“All” subset but one of the entries was defined as “N/A” and the fold-change was lower than one. As such, there were iterations in which this entry, which should pull the score down, was not considered giving rise to a score higher than one. In the third case, TGF-B-1 had four entries in the serum-persistent AF subset. In two of the entries, the protein was defined as overexpressed and in the other two as unchanged. Particularly, one of the “unchanged entries” had a fold-change lower than one and was, as such, the main responsible for pulling the score down. Given that the selection of entries in the bootstrap is random, iterations in which one or both “unchanged entries”, especially the entry with a fold-change lower than one, were not selected had higher scores that contributed to raising the bootstrap score.

Finally, the last case respects to BNP in the plasma-postoperative new-onset AF, which presented four “overexpressed entries”, two of which had especially high fold-changes (4.08 and 3.37), and three “unchanged entries”. Again, the random selection of entries means that not all entries were considered in each iteration, which might lower the coherence parameter and raise the median fold-change parameter. From these four cases, it seems that the bootstrap method presents some limitations regarding the coherence of the findings and outliers and that the simple scoring system performs better in these cases.

Nonetheless, the issues encountered in the bootstrap method might be solved with higher values of p and a larger number of studies focusing on proteins here defined as potential biomarkers for AF; the higher the value of p or the number of entries pertaining to a certain protein, the higher the probability of the final subset in each iteration to have one or even multiple entries of that specific protein. As such, the true scoring value of the protein could be found with more certainty, since the scoring values of each iteration should not differ as much from each other.

Likewise, more studies regarding AF, in its different phases and using all samples, are required, to achieve better results and find the true score of every protein and, subsequently, determine the proteins with the highest biomarker potential, the ones with the highest scores. Additionally, regardless of the subset, every score range started at zero. This also happens because, for each and every protein, no entry or no altered entry was selected in at least one iteration for the final subset because several proteins only had one or few entries in the subset.

5.2 Biomarker Panels

The scoring approach allowed us to identify the proteins with the highest biomarker potential for AF and its different phases and build biomarker panels (**Supplemental Tables 1 – 29**). A variety of proteins/peptides was defined as potential biomarkers for AF in its different phases, but, again, more studies focusing on these proteins/peptides are necessary, since each protein/peptide was studied in AF very few times. Although potential biomarkers for AF were also studied and ranked in the atrial appendages of these patients, the measurement of biological elements in the blood, serum or plasma for diagnostic and prognostic purposes is much more practical in the clinical practice. The measurement of proteins/peptides in the atrial appendages, however, is possible after cardiac surgery, especially to diagnose postoperative new-onset and recurrence of AF. With these considerations in mind, we selected the five potential biomarkers, regardless of the sample, with the highest scores for AF in general ("All" subsets' results) and for each phase of AF (**Tables 5-10**), ignoring the results concerning the atrial appendages subsets for "All", paroxysmal AF, persistent AF and permanent AF as the disease conditions. For the conditions paroxysmal, persistent and permanent AF the chosen proteins were not part of the biomarker panels of the other two conditions, in order to find potential biomarkers specific to each phase and which can act as prognostic indicators. Nonetheless, it is possible that the markers for a certain AF's phase were not defined as potential biomarkers in the other two simply because they were not studied in those conditions and not because they are not altered. Hence, results should be experimentally confirmed. Still, the alteration of the top five biomarkers' levels chosen for each condition seems to agree with AF's pathophysiology.

Interleukin-10 (IL-10), CHI3L1, BNP and NTproBNP measured in the plasma of AF patients and BNP measured in the serum of AF patients were selected as the top potential biomarkers for AF in general. IL-10 inhibits the synthesis of a number of cytokines and is, therefore, involved in a variety of pathways, including the inflammatory response. Although it is normally involved in the inflammatory response provoked by an antigenic stimulus, its overexpression in AF indicates that it might also participate in the non-antigenic inflammatory response. CHI3L1 plays a role in in tissue remodelling and is involved in the inflammatory response. BNP and NTproBNP result from the cleavage of proBNP, which, in turn, is the product of the Natriuretic peptides B. This protein is a cardiac hormone which may function as a paracrine antifibrotic factor in the heart. AF patients are characterized by extensive fibrosis in their atrial tissue, which favours AF's maintenance. As such, BNP and NTproBNP's up-expression probably arises from the up-regulation of the Natriuretic peptides B. As such, the increased levels of IL-10 and CHI3L1 and BNP and NTproBNP likely represent internal response mechanisms to resolve the inflammation process and counterbalance the profibrotic state observed in AF patients, respectively. Furthermore, CHI3L1 might also be partly responsible for the extensive remodelling of the atria.

Interleukin-18 (IL-18), Apelin-12 (APLN12), Vascular cell adhesion protein 1 (VCAM-1), Serum amyloid A-1 protein (SAA1) and Urotensin-2 (U-II) were selected as the top five biomarkers for paroxysmal AF. Il-18 participates in the inflammatory response and VCAM-1 promotes leucocyte migration to the sites of inflammation, which indicates that the up-expression of both proteins emerges in order to solve the pathological lesion that induced inflammation. Contrarily, SAA1 negatively regulates the inflammatory response, which indicates that the absence of SAA1 might be trying to prevent an exacerbation of the inflammatory response in paroxysmal AF patients. Additionally, IL-18 can increase the expression of metalloproteinases [144], which might raise the levels of Matrix metalloproteinase 9 (MMP-9) an other metalloproteinases, culminating in tissue remodelling of the atrial appendages and deposition of fibrotic tissue. APLN12's underexpression in paroxysmal AF suggests that some degree of contractile dysfunction might already exist in paroxysmal AF and that APLN12 might be cleared from the plasma into the atria to contrapose such dysfunction, given its proven ability to improve the recovery of the heart's contractile function in rodents [145] and which might also occur in humans. On the other hand, the down-regulation of this peptide might be, at least in part, responsible for the poor contractility observed in AF patients. U-II is a potent vasoconstrictor and, thus, the up-regulation of this protein might represent an internal response of the organism to the low ejection fraction observed in AF patients due to the contractile dysfunction, which leads to blood accumulation in their atria. Consequently, U-II constricts blood vessels and increments the amount of blood that reaches the heart, to try and normalize the ejection fraction.

Vascular endothelial growth factor receptor 1 (VEGFR-1), ANP, Vascular endothelial growth factor A (VEGF-A), Pentraxin-related protein PTX3 (PTX3) and Hepatocyte growth factor (HGF) represent the top five biomarkers for persistent AF. VEGFR-1 is a negative regulator of VEGF-A, acting in order to limit the amount of free VEGF-A. The first was defined as down-regulated in persistent AF and the second as up-regulated, suggesting that VEGF-A is present in its active bounded form to VEGFR-2 and is highly being used to mediate the inflammatory response of AF individuals. PTX3 is yet another protein defined as overexpressed in persistent AF patients and that also participates in the inflammatory response. ANP is the product of the proANP cleavage, which in turn originates from the Natriuretic peptides A. Hence, it seems likely for ANP's increased concentration to derive from the overexpression of the Natriuretic peptides A, which are secreted in response to the atrial dilation of AF patients, one of the key features of these individuals due to the accumulation of blood in the chambers resultant from poor contraction. HGF is an endothelial specific factor with multiple activities, namely mitogenic, chemoattractant, morphogenic, among others, but possibly its major biological function related to AF is its role as an anti-fibrotic factor [146,147]. Likewise, the raised levels of HGF most likely represent an internal response to counterbalance the deposition or negatively regulate the deposition of fibrotic tissue in the heart of persistent AF subjects.

D-dimer (DD), Tissue-type plasminogen activator (t-PA), Beta-thromboglobulin (B-TG), von Willebrand factor (VWF) and Platelet factor 4 (PF-4) were chosen as the top five biomarkers for permanent AF. DD is involved in fibrin formation, platelet activation and aggregation and in the initiation and maintenance of the process of thrombogenesis. The increased concentration suggests fibrin formation and degradation in permanent AF patients [148]. The second protein is responsible for the conversion of plasminogen to plasmin, controlling, therefore, plasmin-mediated proteolysis and playing a role in tissue remodelling. VWF forms a bridge between sub-endothelial collagen matrix and a platelet receptor complex, to promote platelet adhesion to sites of vascular injury. It also stabilizes and delivers coagulation factor VIII to the site of injury. As such, the overexpression of these three proteins in agreement with the high thrombogenic state seen in AF patients and that, in its turn, may raise the risks of stroke and thromboembolism [140]. Moreover. permanent AF patients have extensive structural remodelling of the heart and, thus, t-PA may be partly responsible or be aggravating the already existing changes. B-TG is the product of Platelet basic protein's cleavage (PBP), a protein which stimulates the formation and secretion of t-PA. The overexpression of B-TG, therefore, suggests an overexpression of PBP, which can help explain the high levels of t-PA encountered in the blood of permanent AF patients and which worsens the thrombogenic process. PF-4 is a protein with chemokine activity, involved in many processes, including platelet activation and degranulation and in the inflammatory response. Hence, PF-4's overexpression seems to represent yet another consequence of the organisms' response to the inflammatory process.

TNFSF11, NTANP, NTproBNP, Growth/differentiation factor 15 (GDF-15) and ANP represent the five potential biomarkers selected for postoperative new-onset AF. TNFSF11 measured in the serum, NTproBNP and CRP measured in the plasma and CRP and TNFSF11 measured in the atrial appendages of patients who developed recurrence of AF after cardioversion were the selected markers for this condition. TNFSF11 plays a role in regulating the inflammatory response [149,150] and the inflammatory factors themselves can contribute to the activation of the axis this protein is involved in [151]. CRP participates in the acute phase response to tissue injury and in the inflammatory response. Therefore, the overexpression of the first in subjects which suffer development of AF after surgery and of both proteins in patients who suffer from AF recurrence occurs in order to counteract tissue inflammation. In addtion, up-regulation of TNFSF11 seems to contribute to the enhancement of gelatinases' activity and a modest decrease of TIMP's expression, which in turn might result in matrix degradation and adverse ventricular remodelling [152]. GDF-15 is a receptor of the TGF-B-1, which in turn is involved in the positive regulation of collagen biosynthetic process and positive regulation of extracellular matrix assembly. Therefore, the up-expression of GDF-15 indicates that levels of TGF-B-1 might also be raised in AF individuals or individuals who are most likely to develop AF. In this manner, the overexpression of TNFSF11 and GDF-15 likely leads to structural remodelling and a higher probability of developing or recurring AF post-surgery. GDF-15 also regulates the apoptotic process, a biological process also seen in AF patients. NTANP is also the product of the cleavage of the Natriuretic peptides A and, as such, its's overexpression in patients which developed AF post-surgery, along with ANP's overexpression, suggests up-regulation of this hormone and some degree of atrial dilation in these patients. NTproBNP's relation to AF was already discussed previously and it seems that this peptide is also a good predictor of AF development or recurrence after surgery.

The proteins here defined as potential biomarkers for AF and each condition, particularly the top five, represent potential biomarkers that can aid physicians, especially when diagnosing AF. As such, the scoring approach here developed depicts a novel method to rank and determine which markers are better to characterize a target disease.

6. Conclusion and Future Work

Our results represent a set of proteins with the highest biomarker potential (highest score) for AF and its different phases and the main biological functions in which they are involved. Alterations in the expression levels of proteins involved in either of these functions seem to agree with AF's pathophysiology and clinical presentation, showing the effectiveness of the developed algorithm. The biomarker panels obtained can be applied to clinical practice for diagnostic and prognostic purposes or even be studied as potential drug targets for AF. Given the high incidence and prevalence rates of AF, the measurement of such markers should be introduced in the clinical analysis routine after the establishment of the threshold values which define normality.

Furthermore, the proposed method is, to our knowledge, the first to reflect the incoherence between studies. Since the scoring function is disease-agnostic it can be applied to datasets concerning other conditions, which should be done in future endeavours. Moreover, other markers, namely genes and metabolites, can also be ranked by the developed system. However, our method is limited by the amount of existing studies. Therefore, experimental tests comparing the levels of the proteins which were highly-scored but have been studied very few times are needed to really understand if they are altered in AF and find their true biomarker potential by, once again, applying the developed method to the dataset containing the new data. Overall, this scoring approach seems to improve the protein's importance harvesting process, which is crucial in a pipeline aimed at predicting potential biomarkers.

References

1. Kannel WB, Abbott RD, Savage DD, McNamara PM. Coronary heart disease and atrial fibrillation: the Framingham Study. Am Heart J. 1983;106:389–96.

2. Feinberg WM, Blackshear JL, Laupacis A, Kronmal R, Hart RG. Prevalence, age distribution, and gender of patients with atrial fibrillation. Analysis and implications. Arch Intern Med. 1995;155:469–73.

3. Kannel W., Wolf P., Benjamin E., Levy D. Prevalence, incidence, prognosis, and predisposing conditions for atrial fibrillation: population-based estimates. Am J Cardiol. 1998;82:2N–9N.

4. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur J Cardiothorac Surg. 2016;50:e1–88.

5. Kearley K, Selwood M, Van den Bruel A, Thompson M, Mant D, Hobbs FR, et al. Triage tests for identifying atrial fibrillation in primary care: a diagnostic accuracy study comparing single-lead ECG and modified BP monitors. BMJ Open. 2014;4:e004565.

6. Fay MR, Fitzmaurice DA, Freedman B. Screening of older patients for atrial fibrillation in general practice: Current evidence and its implications for future practice. Eur J Gen Pract. Informa UK Limited, trading as Taylor & Francis Group; 2017;23:246–53.

7. O'Neal WT, Venkatesh S, Broughton ST, Griffin WF, Soliman EZ. Biomarkers and the prediction of atrial fibrillation: state of the art. Vasc Health Risk Manag. 2016;12:297–303.

8. Nattel S. New ideas about atrial fibrillation 50 years on. Nature. 2002;415:219–26.

9. Chiang C-E, Naditch-Brule L, Murin J, Goethals M, Inoue H, O'Neill J, et al. Distribution and Risk Profile of Paroxysmal, Persistent, and Permanent Atrial Fibrillation in Routine Clinical Practice: Insight From the Real-Life Global Survey Evaluating Patients With Atrial Fibrillation International Registry. Circ Arrhythmia Electrophysiol. 2012;5:632–9.

10. Belluzzi F, Sernesi L, Preti P, Salinaro F, Fonte ML, Perlini S. Prevention of Recurrent Lone Atrial Fibrillation by the Angiotensin-II Converting Enzyme Inhibitor Ramipril in Normotensive Patients. J Am Coll Cardiol. American College of Cardiology Foundation; 2009;53:24–9.

11. Gutierrez C, Blanchard DG. Diagnosis and Treatment of Atrial Fibrillation. Am Fam Physician. 2016;94:442–52.

12. Nattel S, Guasch E, Savelieva I, Cosio FG, Valverde I, Halperin JL, et al. Early management of atrial fibrillation to prevent cardiovascular complications. Eur Heart J. 2014;35:1448–56.

13. Camm AJ, Al-Khatib SM, Calkins H, Halperin JL, Kirchhof P, Lip GYH, et al. A proposal for new clinical concepts in the management of atrial fibrillation. Am Heart J. Mosby, Inc.; 2012;164:292–302.e1.

14. Jahangir A, Lee V, Friedman PA, Trusty JM, Hodge DO, Kopecky SL, et al. Long-Term Progression and Outcomes With Aging in Patients With Lone Atrial Fibrillation: A 30-Year Follow-Up Study. Circulation. 2007;115:3050–6.

15. De Souza AI, Camm AJ. Proteomics of atrial fibrillation. Circ Arrhythmia Electrophysiol. 2012;5:1036–43.

16. Chugh SS, Havmoeller R, Narayanan K, Singh D, Rienstra M, Benjamin EJ, et al. Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. Circulation. 2014;129:837–47.

17. Miyasaka Y, Barnes ME, Gersh BJ, Cha SS, Bailey KR, Abhayaratna WP, et al. Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. Circulation. 2006;114:119–25.

18. Chugh SS, Blackshear JL, Shen WK, Hammill SC, Gersh BJ. Epidemiology and natural history of atrial fibrillation: Clinical implications. J Am Coll Cardiol. Elsevier Masson SAS; 2001;37:371–8.

19. Aguiar C, Macedo ME, Sousa J De, Ferro J, Henriques IL, Rodrigues V, et al. Terapêutica Antitrombótica da Fibrilhação Auricular. Cordenação Nac. para as Doenças Cardiovasc. Lisbon: Textype – Artes Gráficas Lda; 2009. p. 1–24.

20. Ascensão P. Fibrilhação auricular e prevenção do tromboembolismo Estudo numa população de utentes de Centros de Saúde. Rev Port Clin geral. 2006;22:13–24.

21. Ko D, Rahman F, Schnabel RB, Yin X, Benjamin EJ, Christophersen IE. Atrial fibrillation in women: Epidemiology, pathophysiology, presentation, and prognosis. Nat Rev Cardiol. Nature Publishing Group; 2016;13:321–32.

22. Kannel WB, Abbott RD, Savage DD, McNamara PM. Epidemiologic Features of Chronic Atrial Fibrillation. N Engl J Med. 1982;306:1018–22.

23. Wolf P a, Kannel WB, McGee DL, Meeks SL, Bharucha NE, McNamara PM.

Duration of atrial fibrillation and imminence of stroke: the Framingham study. Stroke. 1983;14:664–7.

24. Sentinela M. Divisão de Epidemiologia e Bioestatística da Direcção Geral da Saúde. Um novo olhar sobre a saúde. Lisbon: Direção Geral dos Cuidados de Saúde; 1991.

25. Sentinela M. Divisão de Epidemiologia e Bioestatística da Direcção Geral da Saúde. Um quinto de milhão sob observação. Lisbon: Direção Geral dos Cuidados de Saúde; 1993.

26. Sentinela M. Divisão de Epidemiologia e Bioestatística da Direcção Geral da Saúde. Cinco anos depois. Lisbon: Direção Geral dos Cuidados de Saúde; 1995.

27. Sentinela M. Divisão de Epidemiologia e Bioestatística da Direcção Geral da Saúde. A passo firme. Lisbon: Direção Geral dos Cuidados de Saúde; 1996.

28. Sentinela M. Divisão de Epidemiologia e Bioestatística da Direcção Geral da Saúde. Pela nossa rica saúde. Lisbon: Direção Geral dos Cuidados de Saúde; 1998.

29. Sentinela M. Divisão de Epidemiologia e Bioestatística da Direcção Geral da Saúde. Olhar mais longe. Lisbon: Direção Geral dos Cuidados de Saúde; 1999.

30. Sentinela M. Divisão de Epidemiologia e Bioestatística da Direcção Geral da Saúde. 20 passos em frente. Lisbon: Direção Geral dos Cuidados de Saúde; 2000.

31. Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby J V., et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. JAMA. 2001;285:2370–5.

32. Wolf P a, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke. 1991;22:983–8.

33. Hart RG, Halperin JL. Atrial fibrillation and stroke : concepts and controversies. Stroke. 2001;32:803–8.

34. Kirchhof P. Can we improve outcomes in AF patients by early therapy? BMC Med. 2009;7:72.

35. Mahida S, Ellinor PT. New advances in the genetic basis of atrial fibrillation. J Cardiovasc Electrophysiol. 2012;23:1400–6.

36. Mahida S, Lubitz SA, Rienstra M, Milan DJ, Ellinor PT. Monogenic atrial fibrillation as pathophysiological paradigms. Cardiovasc Res. 2011;89:692–700.

37. Deshmukh A, Barnard J, Sun H, Newton D, Castel L, Pettersson G, et al. Left Atrial Transcriptional Changes Associated With Atrial Fibrillation Susceptibility and

Persistence. Circ Arrhythmia Electrophysiol. 2015;8:32–41.

38. Fox CS. Parental Atrial Fibrillation as a Risk Factor for Atrial Fibrillation in Offspring. JAMA. 2004;291:2851.

39. Heijman J, Voigt N, Nattel S, Dobrev D. Cellular and Molecular Electrophysiology of Atrial Fibrillation Initiation, Maintenance, and Progression. Circ Res. 2014;114:1483–99.

40. Sugiura T, Iwasaka T, Ogawa A, Shiroyama Y, Tsuji H, Onoyama H, et al. Atrial fibrillation in acute myocardial infarction. Am J Cardiol. 1985;56:27–9.

41. Goldberg RJ, Seeley D, Becker RC, Brady P, Chen ZY, Osganian V, et al. Impact of atrial fibrillation on the in-hospital and long-term survival of patients with acute myocardial infarction: a community-wide perspective. Am Hear J. 1990;119:996–1001.

42. Frost L, Mølgaard H, Christiansen EH, Hjortholm K, Paulsen PK, Thomsen PE. Atrial fibrillation and flutter after coronary artery bypass surgery: epidemiology, risk factors and preventive trials. Int J Cardiol. 1992;36:253–61.

43. Lip GY, Beevers DG. ABC of atrial fibrillation. History, epidemiology, and importance of atrial fibrillation. BMJ. 1995;311:1361–3.

44. Murgatroyd FD, Gibson SM, Baiyan X, Nunain SO, Poloniecki JD, Ward DE, et al. Double-Blind Placebo-Controlled Trial of Digoxin in. 1999;

45. Allessie MA, Boyden PA, Camm AJ, Kleber AG, Lab MJ, Legato MJ, et al. Pathophysiology and Prevention of Atrial Fibrillation. Circulation. 2001;103:769–77.

46. Kishore A, Vail A, Majid A, Dawson J, Lees KR, Tyrrell PJ, et al. Detection of atrial fibrillation after ischemic stroke or transient ischemic attack: a systematic review and meta-analysis. Stroke. 2014;45:520–6.

47. Medema MH, van Raaphorst R, Takano E, Breitling R. Computational tools for the synthetic design of biochemical pathways. Nat Rev Microbiol. Nature Publishing Group; 2012;10:191–202.

48. Hjersted JL, Henson MA, Mahadevan R. Genome-scale analysis ofSaccharomyces cerevisiae metabolism and ethanol production in fed-batch culture. Biotechnol Bioeng. 2007;97:1190–204.

49. Lowres N, Neubeck L, Redfern J, Freedman S Ben. Screening to identify unknown atrial fibrillation. A systematic review. Thromb Haemost. 2013;110:213–22.

50. Fitzmaurice DA, Hobbs FDR, Jowett S, Mant J, Murray ET, Holder R, et al. Screening versus routine practice in detection of atrial fibrillation in patients aged 65 or over: cluster randomised controlled trial. Bmj. 2007;335:383–383.

51. Hobbs FR, Taylor CJ, Jan Geersing G, Rutten FH, Brouwer JR. European Primary Care Cardiovascular Society (EPCCS) consensus guidance on stroke prevention in atrial fibrillation (SPAF) in primary care. Eur J Prev Cardiol. 2016;23:460–73.

52. Camm AJ, Lip GYH, De Caterina R, Savelieva I, Atar D, Hohnloser SH, et al. 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: an update of the 2010 ESC Guidelines for the management of atrial fibrillation. Developed with the special contribution of the European Heart Rhythm Association. Eur Heart J. 2012;33:2719–47.

53. Lip GYH, Ramsay SG. Insights from the RCPE UK consensus conference on approaching the comprehensive management of atrial fibrillation. Expert Rev Cardiovasc Ther. 2012;10:697–700.

54. Gaita F, Riccardi R, Gallotti R. Surgical approaches to atrial fibrillation. Card Electrophysiol Rev. 2002;6:401–5.

55. Cox JL, Canavan TE, Schuessler RB, Cain ME, Lindsay BD, Stone C, et al. The surgical treatment of atrial fibrillation. II. Intraoperative electrophysiologic mapping and description of the electrophysiologic basis of atrial flutter and atrial fibrillation. J Thorac Cardiovasc Surg. 1991;101:406–26.

56. Cox JL, Schuessler RB, Lappas DG, Boineau JP. An 8 1/2-year clinical experience with surgery for atrial fibrillation. Ann Surg. 1996;224:267–73.

57. Kottkamp H, Hindricks G, Hammel D, Autschbach R, Mergenthaler J, Borggrefe M, et al. Intraoperative radiofrequency ablation of chronic atrial fibrillation: a left atrial curative approach by elimination of anatomic "anchor" reentrant circuits. J Cardiovasc Electrophysiol. 1999;10:772–80.

58. Sueda T, Nagata H, Orihashi K, Morita S, Okada K, Sueshiro M, et al. Efficacy of a simple left atrial procedure for chronic atrial fibrillation in mitral valve operations. Ann Thorac Surg. 1997;63:1070–5.

59. Roy D, Talajic M, Dorian P, Connolly S, Eisenberg MJ, Green M, et al. Amiodarone to prevent recurrence of atrial fibrillation. Canadian Trial of Atrial Fibrillation Investigators. N Engl J Med. 2000;342:913–20.

60. Nattel S. Experimental evidence for proarrhythmic mechanisms of antiarrhythmic drugs. Cardiovasc Res. 1998;37:567–77.

61. Kirchhof C, Chorro F, Scheffer GJ, Brugada J, Konings K, Zetelaki Z, et al. Regional entrainment of atrial fibrillation studied by high-resolution mapping in open-chest dogs. Circulation. 1993;88:736–49.

62. Haïssaguerre M, Jaïs P, Shah DC, Takahashi A, Hocini M, Quiniou G, et al. Spontaneous Initiation of Atrial Fibrillation by Ectopic Beats Originating in the Pulmonary Veins. N Engl J Med. 1998;339:659–66.

63. Shah AN, Mittal S, Sichrovsky TC, Cotiga D, Arshad A, Maleki K, et al. Long-term outcome following successful pulmonary vein isolation: pattern and prediction of very late recurrence. J Cardiovasc Electrophysiol. 2008;19:661–7.

64. Prasad SM, Maniar HS, Camillo CJ, Schuessler RB, Boineau JP, Sundt TM, et al. The Cox maze III procedure for atrial fibrillation: long-term efficacy in patients undergoing lone versus concomitant procedures. J Thorac Cardiovasc Surg. 2003;126:1822–8.

65. de Vos CB, Pisters R, Nieuwlaat R, Prins MH, Tieleman RG, Coelen R-JS, et al. Progression from paroxysmal to persistent atrial fibrillation clinical correlates and prognosis. J Am Coll Cardiol. Elsevier Inc.; 2010;55:725–31.

66. Silverman ME. From rebellious palpitations to the discovery of auricular fibrillation: Contributions of Mackenzie, Lewis and Einthoven. Am J Cardiol. 1994;73:384–9.

67. Yue L, Feng J, Gaspo R, Li GR, Wang Z, Nattel S. Ionic remodeling underlying action potential changes in a canine model of atrial fibrillation. Circ Res. 1997;81:512–25.

68. Schotten U, Duytschaever M, Ausma J, Eijsbouts S, Neuberger HR, Allessie M. Electrical and contractile remodeling during the first days of atrial fibrillation go hand in hand. Circulation. 2003;107:1433–9.

69. De Souza AI, Cardin S, Wait R, Chung YL, Vijayakumar M, Maguy A, et al. Proteomic and metabolomic analysis of atrial profibrillatory remodelling in congestive heart failure. J Mol Cell Cardiol. Elsevier Ltd; 2010;49:851–63.

70. Ausma J, Wijffels M, Thoné F, Wouters L, Allessie M, Borgers M. Structural changes of atrial myocardium due to sustained atrial fibrillation in the goat. Circulation. 1997;96:3157–63.

71. Frustaci A, Caldarulo M, Buffon A, Bellocci F, Fenici R, Melina D. Cardiac biopsy in patients with “primary” atrial fibrillation; Histologic evidence of occult myocardial diseases. Chest. The American College of Chest Physicians; 1991;100:303–6.

72. Heijman J, Voigt N, Nattel S, Dobrev D. Cellular and molecular electrophysiology of atrial fibrillation initiation, maintenance, and progression. Circ Res. 2014;114:1483–99.

73. Nattel S, Wakili R, Dobrev D, Voigt N, Kääb S. Recent advances in the molecular pathophysiology of atrial fibrillation. J Clin Invest. 2011;121:2955–68.

74. Ng J, Villuendas R, Cokic I, Schliamser JE, Gordon D, Koduri H, et al. Autonomic remodeling in the left atrium and pulmonary veins in heart failure creation of a dynamic substrate for atrial fibrillation. Circ Arrhythmia Electrophysiol. 2011;4:388–96.

75. Nishida K, Qi XY, Wakili R, Comtois P, Chartier D, Harada M, et al. Mechanisms of atrial tachyarrhythmias associated with coronary artery occlusion in a chronic canine model. Circulation. 2011;123:137–46.

76. Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, et al. The case for early detection. Nat Rev Cancer. 2003;3:243–52.

77. Koura T, Hara M, Takeuchi S, Ota K, Okada Y, Miyoshi S, et al. Anisotropic conduction properties in canine atria analyzed by high-resolution optical mapping: Preferential direction of conduction block changes from longitudinal to transverse with increasing age. Circulation. 2002;105:2092–8.

78. P C, R A. Cardiac Arrhythmias: The Role of Triggered Activity and Other Mechanisms. New York, NY Futur. 1988;

79. MacLennan DH, Chen SRW. Store overload-induced Ca2+ release as a triggering mechanism for CPVT and MH episodes caused by mutations in RYR and CASQ genes. J Physiol. 2009;587:3113–5.

80. Zellerhoff S, Pistulli R, Mönnig G, Hinterseer M, Beckmann BM, Köbe J, et al. Atrial arrhythmias in long-QT syndrome under daily life conditions: A nested case control study. J Cardiovasc Electrophysiol. 2009;20:401–7.

81. Lemoine MD, Duverger JE, Naud P, Chartier D, Qi XY, Comtois P, et al. Arrhythmogenic left atrial cellular electrophysiology in a murine genetic long QT syndrome model. Cardiovasc Res. 2011;92:67–74.

82. Nattel S, Dobrev D. Electrophysiological and molecular mechanisms of paroxysmal atrial fibrillation. Nat Rev Cardiol. 2016;13:575–90.

83. Dobrev D, Nattel S. New antiarrhythmic drugs for treatment of atrial fibrillation. Lancet. Elsevier Ltd; 2010;375:1212–23.

84. Nattel S, Burstein B, Dobrev D. Atrial remodeling and atrial fibrillation: mechanisms and implications. Circ Arrhythm Electrophysiol. 2008;1:62–73.

85. Verheule S, Wilson E, Iv TE, Shanbhag S, Olgin J. Alterations in Atrial Electrophysiology and Tissue Structure Mitral Regurgitation. 2003;2615–22.

86. John B, Stiles MK, Kuklik P, Chandy ST, Young GD, MacKenzie L, et al.

Electrical remodelling of the left and right atria due to rheumatic mitral stenosis. Eur Heart J. 2008;29:2234–43.

87. Rajesh T, Park H-Y, Song E, Sung C, Park S-H, Lee J-H, et al. A new flow path design for multidimensional protein identification technology using nano-liquid chromatography electrospray ionization mass spectrometry. Korean J Chem Eng. 2013;30:417–21.

88. Strimbu K, Tavel J a. What are biomarkers? Curr Opin HIV AIDS. 2010;5:463–6.

89. Gerszten RE, Wang TJ. The search for new cardiovascular biomarkers. Nature. 2008;451:949–52.

90. O'Gara PT, Kushner FG, Ascheim DD, Casey DE, Chung MK, de Lemos JA, et al. 2013 ACCF/AHA Guideline for the Management of ST-Elevation Myocardial Infarction: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation. 2013;127:e362–425.

91. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, et al. 2013 ACCF/AHA guideline for the management of heart failure: A report of the American college of cardiology foundation/american heart association task force on practice guidelines. J Am Coll Cardiol. Elsevier; 2013;62:e147–239.

92. Guideline AHA, Non PW, Syndromes SAC, Guidelines P, Angiography C, With C, et al. Erratum: 2014 ACC/AHA guideline for the management of patients with non-ST-elevation acute coronary syndromes: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Circulation (2014) 130 (e344-e426)). Circulation. 2014;130:e433–4.

93. Smith JG, Newton-Cheh C, Almgren P, Struck J, Morgenthaler NG, Bergmann A, et al. Assessment of conventional cardiovascular risk factors and multiple biomarkers for the prediction of incident heart failure and atrial fibrillation. J Am Coll Cardiol. 2010;56:1712–9.

94. Fu R, Wu S, Wu P, Qiu J. A Study of blood soluble P-selectin, fibrinogen, and von Willebrand factor levels in idiopathic and lone atrial fibrillation. Europace. 2011;13:31–6.

95. Begg GA, Lip GYH, Plein S, Tayebjee MH. Circulating biomarkers of fibrosis and cardioversion of atrial fibrillation: A prospective, controlled cohort study. Clin Biochem. Elsevier B.V.; 2017;50:11–5.

96. Chang K-W, Hsu JC, Toomu A, Fox S, Maisel AS. Clinical Applications of Biomarkers in Atrial Fibrillation. Am J Med. Elsevier Inc.; 2017;

97. Rienstra M, Yin X, Larson MG, Fontes JD, Magnani JW, McManus DD, et al. Relation between soluble ST2, growth differentiation factor-15, and high-sensitivity troponin i and incident atrial fibrillation. Am Heart J. Mosby, Inc.; 2014;167:109–115.e2.

98. Filion KB, Agarwal SK, Ballantyne CM, Eberg M, Hoogeveen RC, Huxley RR, et al. High-sensitivity cardiac troponin T and the risk of incident atrial fibrillation: The Atherosclerosis Risk in Communities (ARIC) study. Am Heart J. Elsevier Inc.; 2015;169:31–38.e3.

99. Hussein AA, Bartz TM, Gottdiener JS, Sotoodehnia N, Heckbert SR, Lloyd-Jones D, et al. Serial measures of cardiac troponin T levels by a highly sensitive assay and incident atrial fibrillation in a prospective cohort of ambulatory older adults. Hear Rhythm. Elsevier; 2015;12:879–85.

100. Masson S, Wu JHY, Simon C, Barlera S, Marchioli R, Mariani J, et al. Circulating cardiac biomarkers and postoperative atrial fibrillation in the OPERA trial. Eur J Clin Invest. 2015;45:170–8.

101. Narducci ML, Pelargonio G, Rio T, Leo M, Di Monaco A, Musaico F, et al. Predictors of postoperative atrial fibrillation in patients with coronary artery disease undergoing cardiopulmonary bypass: A possible role for myocardial ischemia and atrial inflammation. J Cardiothorac Vasc Anesth. Elsevier; 2014;28:512–9.

102. Leal JC, Petrucci O, Godoy MF, Braile DM. Perioperative serum troponin i levels are associated with higher risk for atrial fibrillation in patients undergoing coronary artery bypass graft surgery. Interact Cardiovasc Thorac Surg. 2012;14:22–5.

103. Knayzer B, Abramov D, Natalia B, Tovbin D, Ganiel A, Katz A. Atrial fibrillation and plasma troponin I elevation after cardiac surgery: Relation to inflammation-associated parameters. J Card Surg. 2007;22:117–23.

104. Sörensen NA, Shah AS, Ojeda FM, Peitsmeyer P, Zeller T, Keller T, et al. High-sensitivity troponin and novel biomarkers for the early diagnosis of non-ST-segment elevation myocardial infarction in patients with atrial fibrillation. Eur Hear J Acute Cardiovasc Care. 2016;5:419–27.

105. Liebetrau C, Weber M, Tzikas S, Palapies L, Möllmann H, Pioro G, et al. Identification of acute myocardial infarction in patients with atrial fibrillation and chest pain with a contemporary sensitive troponin I assay. BMC Med. BMC Medicine; 2015;13:169.

106. Hijazi Z, Oldgren J, Andersson U, Connolly SJ, Ezekowitz MD, Hohnloser SH, et al. Importance of persistent elevation of cardiac biomarkers in atrial fibrillation: a RE-

LY substudy. Heart. 2014;100:1193–200.

107. Wang TJ, Larson MG, Levy D, Benjamin EJ, Leip EP, Omland T, et al. Plasma Natriuretic Peptide Levels and the Risk of Cardiovascular Events and Death. N Engl J Med. 2004;350:655–63.

108. Matsuura H, Murakami T, Hina K, Yamamoto K, Kawamura H, Sogo T, et al. Association of elevated plasma B-type natriuretic peptide levels with paroxysmal atrial fibrillation in patients with nonobstructive hypertrophic cardiomyopathy. Clin Biochem. 2008;41:134–9.

109. Poste G. Bring on the biomarkers. Nature. 2011;469:156–7.

110. Baber U, Howard VJ, Halperin JL, Soliman EZ, Zhang X, McClellan W, et al. Association of chronic kidney disease with atrial fibrillation among adults in the United States REasons for Geographic and Racial Differences in Stroke (REGARDS) study. Circ Arrhythmia Electrophysiol. 2011;4:26–32.

111. Modrego J, Maroto L, Tamargo J, Azcona L, Mateos-Cáceres P, Segura A, et al. Comparative expression of proteins in left and right atrial appendages from patients with mitral valve disease at sinus rhythm and atrial fibrillation. J Cardiovasc Electrophysiol. 2010;21:859–68.

112. Prensner JR, Chinnaiyan AM, Srivastava S. Systematic, evidence-based discovery of biomarkers at the NCI. Clin Exp Metastasis. 2012;29:645–52.

113. Bravo A, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI. A knowledge-driven approach to extract disease-related biomarkers from the literature. Biomed Res Int. 2014;2014.

114. Heo GE, Kang KY, Song M, Lee J-H. Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. BMC Bioinformatics. 2017;18:251.

115. García-Sancho M. From metaphor to practices: The introduction of “information engineers” into the first DNA sequence database. Hist Philos Life Sci. 2011;33:71–104.

116. Bartlett A, Penders B, Lewis J. Bioinformatics: indispensable, yet hidden in plain sight? BMC Bioinformatics. BMC Bioinformatics; 2017;18:311.

117. van Baren-Nawrocka J. The bioinformatics of genetic origins: how identities become embedded in the tools and practices of bioinformatics. Life Sci Soc Policy. 2013;9:7.

118. Lewis J, Bartlett A. Inscribing a discipline: tensions in the field of bioinformatics. New Genet Soc. 2013;32:243–63.

119. Xu D, Zhang M, Xie Y, Wang F, Chen M, Zhu KQ, et al. DTMiner: identification

of potential disease targets through biomedical literature mining. Bioinformatics. 2016;32:3619–26.

120. Wishart D, Tzu D, Knox C. HMDB: the Human Metabolome Database [Internet]. Nucleic Acids Res. 2007 [cited 2017 Nov 28]. Available from: http://www.hmdb.ca/diseases

121. Solutions EK. GOBIOM - Global Online Biomarker Database [Internet]. 2017 [cited 2017 Nov 28]. Available from: https://gobiomdb.com/login.jsp#overview

122. Liu R-L, Shih C-C. Identification of highly related references about gene-disease association. BMC Bioinformatics. 2014;15:286.

123. Automated vs Manual Literature Curation. Elsevier. 2014;

124. Sernadela P, Oliveira JL. A semantic-based workflow for biomedical literature annotation. Database (Oxford). 2017;846–60.

125. Hirschman J, Berardini TZ, Drabkin HJ, Howe D. A MOD(ern) perspective on literature curation. Mol Genet Genomics. 2010;283:415–25.

126. Lourenço A, Carreira R, Carneiro S, Maia P, Glez-Peña D, Fdez-Riverola F, et al. @Note: A workbench for Biomedical Text Mining. J Biomed Inform. Elsevier Inc.; 2009;42:710–20.

127. Chesler EJ, Haendel M. Text mining versus manual curation. Int Rev Neurobiol Bioinforma Behav Part 2. 1st ed. Academic Press; 2012. p. 12–3.

128. Zweigenbaum P, Demner-Fushman D, Yu H, KB C. Frontiers of biomedical text mining: current progress. Br Bioinform. 2007;8:358–75.

129. Liu Y, Navathe SB, Civera J, Dasigi V, Ram A, Ciliax BJ, et al. Text mining biomedical literature for discovering gene-to-gene relationships: A comparative study of algorithms. IEEE/ACM Trans Comput Biol Bioinforma. 2005;2:62–75.

130. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. World Wide Web Internet Web Inf Syst. 1998;54:1–17.

131. GRIB/IMIM/UPF IBIG. Gene-disease association data retrieved from DisGeNET v5.0 [Internet]. [cited 2017 Jan 24]. Available from: http://www.disgenet.org/

132. Ernst M, Du Y, Warsow G, Hamed M, Endlich N, Endlich K, et al. FocusHeuristics – expression-data-driven network optimization and disease gene prediction. Sci Rep. Nature Publishing Group; 2017;7:42638.

133. Yu F, Yang Z, Hu X, Sun Y, Lin H, Wang J. Protein complex detection in PPI networks based on data integration and supervised learning method. BMC Bioinformatics. BioMed Central Ltd; 2015;16:S3.

134. Mihm MJ, Yu F, Carnes CA, Reiser PJ, Mccarthy PM, Wagoner DR Van, et al. During Human Atrial Fibrillation. 2015;174–81.

135. Ausma J, Coumans WA, Duimel H, Van Der Vusse GJ, Allessie MA, Borgers M. Atrial high energy phosphate content and mitochondrial enzyme activity during chronic atrial fibrillation. Cardiovasc Res. 2000;47:788–96.

136. Dzeshka MS, Lip GYH, Snezhitskiy V, Shantsila E. Cardiac Fibrosis in Patients With Atrial Fibrillation: Mechanisms and Clinical Implications. J Am Coll Cardiol. 2015;66:943–59.

137. Guo Y, Lip GYH, Apostolakis S. Inflammation in atrial fibrillation. J Am Coll Cardiol. Elsevier Inc.; 2012;60:2263–70.

138. Savelieva I, Kakouros N, Kourliouros A, Camm AJ. Upstream therapies for management of atrial fibrillation: Review of clinical evidence and implications for European Society of Cardiology guidelines. Part I: Primary prevention. Europace. 2011;13:308–28.

139. Hu YF, Chen YJ, Lin YJ, Chen SA. Inflammation and the pathogenesis of atrial fibrillation. Nat Rev Cardiol. Nature Publishing Group; 2015;12:230–43.

140. Lip GY, Lowe GD. Fibrin D-dimer: a useful clinical marker of thrombogenesis? Clin Sci (Lond). 1995;89:205–14.

141. Van Wagoner DR. Molecular basis of atrial fibrillation: A dream or a reality? J Cardiovasc Electrophysiol. 2003;14:667–9.

142. Korantzopoulos P, Kolettis T, Siogas K, Goudevenos J. Atrial fibrillation and electrical remodeling: the potential role of inflammation and oxidative stress. Med Sci Monit. 2003;9:RA225-9.

143. Van Wagoner DR. Electrophysiological remodeling in human atrial fibrillation. Pacing Clin Electrophysiol. 2003;26:1572–5.

144. Ishida Y, Migita K, Izumi Y, Nakao K, Ida H, Kawakami A, et al. The role of IL-18 in the modulation of matrix metalloproteinases and migration of human natural killer (NK) cells. FEBS Lett. 2004;569:156–60.

145. Pisarenko OI, Serebryakova LI, Pelogeykina YA, Studneva IM, Khatri DN, Tskitishvili O V., et al. In vivo reduction of reperfusion injury to the heart with apelin-12 peptide in rats. Bull Exp Biol Med. 2011;152:79–82.

146. Okunishi K, Dohi M, Nakagome K, Tanaka R, Mizuno S, Matsumoto K, et al. A Novel Role of Hepatocyte Growth Factor as an Immune Regulator through Suppressing Dendritic Cell Function. J Immunol. 2005;175:4745–53.

147. Ishikawa H, Jo J, Tabata Y. Liver Anti-Fibrosis Therapy with Mesenchymal Stem Cells Secreting Hepatocyte Growth Factor Liver Anti-Fibrosis Therapy with Mesenchymal Stem. 2017;5063:37–41.

148. Lip GYH, Lip PL, Zarifis J, Watson RDS, Bareford D, Lowe GDO, et al. Fibrin D-Dimer and Beta Thromboglobulin as Markers of Thrombogenesis and Platelet Activation in Atrial Fibrillation: Effects of Introducing Ultra Low-Dose Warfarin and Aspirin. Circulation. 1996;94:425–31.

149. Yun TJ, Tallquist MD, Aicher A, Rafferty KL, Marshall AJ, Moon JJ, et al. Osteoprotegerin, a Crucial Regulator of Bone Metabolism, Also Regulates B Cell Development and Function. J Immunol. 2001;166:1482–91.

150. Mosheimer BA, Kaneider NC, Feistritzer C, Sturn DH, Wiedermann CJ. Expression and function of RANK in human monocyte chemotaxis. Arthritis Rheum. 2004;50:2309–16.

151. Ho T-Y, Santora K, Chen JC, Frankshun A-L, Bagnell CA. Effects of relaxin and estrogens on bone remodeling markers, receptor activator of NF-kB ligand (RANKL) and osteoprotegerin (OPG), in rat adjuvant-induced arthritis. Bone. Elsevier Inc.; 2011;48:1346–53.

152. Ueland T, Yndestad A, Øie E, Florholmen G, Halvorsen B, Frøland SS, et al. Dysregulated osteoprotegerin/RANK ligand/RANK axis in clinical and experimental heart failure. Circulation. 2005;111:2461–8.

Appendix

Supplemental Table 1 – Atrial Appendages-"All" subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
P11217	Glycogen phosphorylase, muscle form	PYGM	PYGM	15.33	Down	1
Q01995	Transgelin	TAGLN	TAGLN	6.39	Down	1
P15531	Nucleoside diphosphate kinase A	NDKA	NME1	4.45	Down	1
P06756	Integrin alpha-V	ITGAV	ITGAV	3.81	Up	6
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	3.45	Up	4
P01137	Transforming growth factor beta-1	TGF-B-1	TGFB1	3.25	Up	8
Q9Y6Q6	Tumor necrosis factor receptor superfamily member 11A	TNFRSF11A	TNFRSF11A	3.12	Up	4
P27797	Calreticulin	CALR	CALR	3.11	Up	6
Q99807	5-demethoxyubiquinone hydroxylase, mitochondrial	DMQH	COQ7	2.86	Up	1
Q86TX2	Acyl-coenzyme A thioesterase 1	ACOT1	ACOT1	2.77	Down	1
P23142	Fibulin-1	FIBL-1	FBLN1	2.52	Down	2
Q96CX2	BTB/POZ domain-containing protein KCTD12	KCTD12	KCTD12	2.49	Up	1
P09936	Ubiquitin carboxyl-terminal hydrolase isozyme L1	UCH-L1	UCHL1	2.3	Down	1
P31150	Rab GDP dissociation inhibitor alpha	RabGDIA	GDI1	2.12	Down	1
P02765	Alpha-2-HS-glycoprotein	AHSG	AHSG	2.12	Up	1
P14384	Carboxypeptidase M	CPM	CPM	2.08	Down	1
O00300	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	TNFRSF11B	1.87	Up	4
P02741	C-reactive protein	CRP	CRP	1.53	Up	2

Supplemental Table 2 – Whole blood-“All” subset’s potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
-	N-terminal pro-Brain Natriuretic peptide	NTproBNP	NPPB	2.11	Up	3
-	Mid-region pro-Atrial Natriuretic peptide	MRproANP	NPPA	1.39	Up	3
P16671	Platelet glycoprotein 4	PG-4	CD36	1.21	Down	1
P01034	Cystatin-C	CST-C	CST3	1.13	Up	3

Supplemental Table 3 - Plasma-“All” subset’s potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Protein	Gene	Score	Direction	Number of Entries
P22301	Interleukin-10	IL-10	IL10	5.86	Up	2
P36222	Chitinase-3-like protein 1	CHI3L1	CHI3L1	2.54	Up	2
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	2.33	Up	10
-	Brain Natriuretic Peptide	BNP	NPPB	2.22	Up	17
-	D-dimer	DD	-	2.17	Up	1
P43235	Cathepsin K	CTSK	CTSK	2.15	Up	1
-	N-terminal pro-Atrial Natriuretic Peptide	NTproANP	NPPA	1.83	Up	1
P45379	Troponin T, cardiac muscle	TnTc	TNNT2	1.70	Up	1
-	N-terminal Atrial Natriuretic Peptide	NTANP	NPPA	1.58	Up	1
P19440	Glutathione hydrolase 1 proenzyme	GH1	GGT1	1.49	Up	1
P05362	Intercellular adhesion molecule 1	ICAM1	ICAM1	1.44	Up	2
-	Atrial Natriuretic Peptide	ANP	NPPA	1.34	Up	9
P01033	Metalloproteinase inhibitor 1	TIMP-1	TIMP1	1.30	Up	2
-	Beta-thromboglobulin	B-TG	PPBP	1.28	Up	3
P23142	Fibulin-1	FIBL-1	FBLN1	1.27	Up	1
P14780	Matrix metalloproteinase-9	MMP-9	MMP9	1.16	Up	1

Supplemental Table 4 - Serum-"All" subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Protein	Gene	Score	Direction	Number of Entries
-	Brain Natriuretic Peptide	BNP	NPPB	2.34	Up	2
-	Atrial Natriuretic Peptide	ANP	NPPA	1.99	Up	4
-	Relaxin	RLX	RLN2	1.51	Up	2
P01584	Interleukin-1 beta	IL-1B	IL1B	1.31	Up	1
P01374	Lymphotoxin-alpha	TNF-B	LTA	1.22	Up	2
P14780	Matrix metalloproteinase-9	MMP-9	MMP9	1.20	Up	5
P69905	Hemoglobin subunit alpha	HBA1	HBA1	1.08	Up	3
Q14116	Interleukin-18	IL-18	IL18	1.02	Up	7

Supplemental Table 5 - Atrial appendages-paroxysmal AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	3.7	Up	2
Q9Y6Q6	Tumor necrosis factor receptor superfamily member 11A	TNFRSF11A	TNFRSF11A	3.35	Up	2
P06756	Integrin alpha-V	ITGAV	ITGAV	3.22	Up	2
O00300	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	TNFRSF11B	1.82	Up	2
P01137	Transforming growth factor beta-1	TGF-B-1	TGFB1	1.77	UP	2
P27797	Calreticulin	CALR	CALR	1.49	UP	2
P02741	C-reactive protein	CRP	CRP	1.44	UP	1

Supplemental Table 6 – Plasma-paroxysmal AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
-	Brain Natriuretic Peptide	BNP	NPPB	2.66	Up	3
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	2.49	Up	3
-	Atrial Natriuretic Peptide	ANP	NPPA	2.07	Up	5
-	N-terminal pro-Atrial Natriuretic Peptide	NTproANP	NPPA	1.95	Up	1
-	Apelin-12	APLN12	-	1.63	Down	1
P19320	Vascular cell adhesion protein 1	VCAM-1	VCAM1	1.55	Up	1
O95399	Urotensin-2	U-II	UTS2	1.44	Up	1

Supplemental Table 7 – Serum-paroxysmal AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
Q14116	Interleukin-18	IL-18	IL18	1.78	Up	1
-	Brain natriuretic peptide	BNP	NPPB	1.72	Up	1
P0DJI8	Serum amyloid A-1 protein	SAA1	SAA1	1.45	Up	1
-	Relaxin	RLX	RLN2	1.44	Up	1
Q99988	Growth/differentiation factor 15	GDF-15	GDF15	1.38	Up	1
Q99727	Metalloproteinase inhibitor 4	TIMP-4	TIMP4	1.32	Up	1
P01375	Tumor necrosis factor	TNF-A	TNF	1.31	Up	2
P01374	Lymphotoxin-alpha	TNF-B	LTA	1.21	Up	1
-	Neuregulin-1	NRG1	NRG1	1.19	Up	1
Q9HD89	Resistin	RETN	RETN	1.14	Up	1
P16035	Metalloproteinase inhibitor 2	TIMP-2	TIMP2	1.12	Up	1
P69905	Hemoglobin subunit alpha	HBA1	HBA1	1.08	Up	1

Supplemental Table 8 – Atrial appendages-persistent AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Protein	Gene	Score	Direction	Number of Entries
P62937	Peptidyl-prolyl cis-trans isomerase A	PPIaseA	PPIA	5.50	Up	1
P49748	Very long-chain specific acyl-CoA dehydrogenase, mitochondrial	VLCAD	ACADVL	4.50	Up	1
P27797	Calreticulin	CALR	CALR	4.23	Up	4
P06756	Integrin alpha-V	ITGAV	ITGAV	3.94	Up	4
P01137	Transforming growth factor beta-1	TGF-B-1	TGFB1	3.71	Up	6
P00441	Superoxide dismutase Cu-Zn	SOD1	SOD1	3.50	Up	1
P08590	Myosin light chain 3	MYL3	MYL3	3.30	Up	1
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	3.28	Up	2
P30048	Thioredoxin-dependent peroxide reductase, mitochondrial	TDPRDX	PRDX3	3.20	Down	1
Q9Y6Q6	Tumor necrosis factor receptor superfamily member 11A	TNFRSF11A	TNFRSF11A	2.97	Up	2
P12821	Angiotensin-converting enzyme	ACE	ACE	2.94	Up	1
Q06830	Peroxiredoxin-1	PRDX1	PRDX1	2.90	Up	1
Q9P0J0	NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 13	NDUFA13	NDUFA13	2.80	Up	1
P23528	Cofilin-1	CFL1	CFL1	2.60	Up	1
P30086	Phosphatidylethanolamine-binding protein 1	PEBP-1	PEBP1	2.60	Up	1
O95299	NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 10, mitochondrial	NDUFA10	NDUFA10	2.30	Down	1
P19429	Troponin I, cardiac muscle	TnIc	TNNI3	2.30	Up	1
Q13011	Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase, mitochondrial	DDDCoAI	ECH1	2.20	Down	1
P04264	Keratin, type II cytoskeletal 1	K1	KRT1	2.10	Down	1
P45880	Voltage-dependent anion-selective channel protein 2	VDAC-2	VDAC2	2.10	Up	1
Q15796	Mothers against decapentaplegic homolog 2	MADH2	SMAD2	2.00	Up	2

Supplemental Table 9 – Atrial appendages-persistent AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset (continued).

UNIPROT Code	Full Name	Protein	Gene	Score	Direction	Number of Entries
O00300	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	TNFRSF11B	1.94	Up	2
P02741	C-reactive protein	CRP	CRP	1.62	Up	1
P35609	Alpha-actinin-2	ACTN2	ACTN2	1.61	Up	2

Supplemental Table 10 – Whole blood-persistent AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
P02741	C-reactive protein	CRP	CRP	2.17	Up	3
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	1.95	Up	1
-	Mid-region pro-Atrial Natriuretic Peptide	MRproANP	NPPA	1.39	Up	1
P16109	P-selectin	SELP	SELP	1.38	Up	1
P01034	Cystatin-C	CST-C	CST3	1.13	Up	1

Supplemental Table 11 – Serum-persistent AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
-	Atrial Natriuretic Peptide	ANP	NPPA	3.56	Up	1
-	Brain Natriuretic Peptide	BNP	NPPB	2.96	Up	1
P26022	Pentraxin-related protein PTX3	PTX3	PTX3	1.96	Up	4
P14210	Hepatocyte growth factor	HGF	HGF	1.60	Up	3
-	Relaxin	RLX	RLN2	1.58	Up	1
P14780	Matrix metalloproteinase-9	MMP-9	MMP9	1.56	Up	1
P25445	Tumor necrosis factor receptor superfamily member 6	TNFRSF6	FAS	1.56	Up	3
P01374	Lymphotoxin-alpha	TNF-B	LTA	1.24	Up	1
Q9HD89	Resistin	RETN	RETN	1.16	Up	1
P16581	E-selectin	SELE	SELE	1.16	Up	1
P08253	72 kDa type IV collagenase	MMP-2	MMP2	1.14	Down	1
P69905	Hemoglobin subunit alpha	HBA1	HBA1	1.08	Up	1

Supplemental Table 12 – Atrial appendages-permanent AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
P11177	Pyruvate dehydrogenase E1 component subunit beta, mitochondrial	PDHE1-B	PDHB	3.87	Up	2
P06753	Tropomyosin alpha-3 chain	TPM3	TPM3	2.77	Up	2
P12829	Myosin light chain 4	MYL4	MYL4	2.56	Up	2
P07951	Tropomyosin beta chain	TMSB	TPM2	2.05	Up	2
P30084	Enoyl-CoA hydratase, mitochondrial	ECoAh	ECHS1	1.77	Down	2
P10809	60 kDa heat shock protein, mitochondrial	Hsp60	HSPD1	1.75	Up	8
P68032	Actin, alpha cardiac muscle 1	ACTC1	ACTC1	1.51	Up	6
P04792	Heat shock protein beta-1	HspB1	HSPB1	1.26	Up	2
P06732	Creatine kinase M-type	M-CK	CKM	1.07	Up	4

Supplemental Table 13 – Whole blood-permanent AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	3.51	Up	1
-	D-dimer	DD	-	2.71	Up	1
P00750	Tissue-type plasminogen activator	t-PA	PLAT	2.08	Up	1
-	Mid-region pro-Atrial Natriuretic Peptide	MRproANP	NPPA	1.78	Up	1
P04275	von Willebrand factor	VWF	VWF	1.76	Up	1
P02741	C-reactive protein	CRP	CRP	1.72	Up	1
P05121	Plasminogen activator inhibitor 1	PAI	SERPINE1	1.63	Up	1
P07204	Thrombomodulin	TM	THBD	1.46	Up	1
P00750	Tissue-type plasminogen activator	CST-C	CST3	1.18	Up	1

Supplemental Table 14 – Plasma-permanent AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
P36222	Chitinase-3-like protein 1	CHI3L1	CHI3L1	3.37	Up	1
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	2.16	Up	1
-	Beta-thromboglobulin	B-TG	PPBP	1.96	Up	1
P02776	Platelet factor 4	PF-4	PF4	1.70	Up	1
P02741	C-reactive protein	CRP	CRP	1.27	Up	4
P13726	Tissue Factor	TF	F3	1.18	Up	2

Supplemental Table 15 – Plasma-postoperative new-onset AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
Q99988	Growth/differentiation factor 15	GDF-15	GDF15	1.56	Up	1
-	Atrial Natriuretic Peptide	ANP	NPPA	1.44	Up	6
-	N-terminal Atrial Natriuretic Peptide	NTANP	NPPA	1.34	Up	2
P07204	Thrombomodulin	TM	THBD	1.25	Up	1

Supplemental Table 16 – Serum-postoperative new-onset AF subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	2.71	Up	1
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	1.60	Up	2
O00300	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	TNFRSF11B	1.15	Up	1
P19429	Troponin I, cardiac muscle	TnIc	TNNI3	1.08	Up	2

Supplemental Table 17 – Plasma-postoperative AF recurrence subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	1.92	Up	2
P02741	C-reactive protein	CRP	CRP	1.66	Up	7
P05231	Interleukin-6	IL-6	IL6	1.40	Up	1
-	mid-regional pro-adrenomedullin	MRproAD	ADM	1.37	Up	1
-	pro-Atrial Natriuretic Peptide	proANP	NPPA	1.36	Up	1
Q9ULZ1	Apelin	APLN	APLN	1.23	Down	1
P48061	Stromal cell-derived factor 1	SDF-1	CXCL12	1.20	Up	1
-	N-terminal pro-Atrial Natriuretic Peptide	NTproANP	NPPA	1.07	Up	1

Supplemental Table 18 – Serum-postoperative AF recurrence subset's potential biomarker proteins and respective UNIPROT code, full name, abbreviations, gene, score, computation direction and number of entries in the subset.

UNIPROT Code	Full Name	Abbreviation	Gene	Score	Direction	Number of Entries
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	2.36	Up	1
-	N-terminal pro-Brain Natriuretic Peptide	GH1	NPPB	1.50	Up	1
P16035	Metalloproteinase inhibitor 2	NTproBNP	TIMP2	1.34	Up	2
O00300	Tumor necrosis factor receptor superfamily member 11B	TIMP-2	TNFRSF11B	1.26	Up	1
P08253	72 kDa type IV collagenase	TNFRSF11B	MMP2	1.20	Up	1
O14788	Tumor necrosis factor ligand superfamily member 11	MMP-2	TNFSF11	1.11	Up	1

Supplemental Table 19 – Atrial appendages-“All” subset’s potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
P06756	Integrin alpha-V	ITGAV	ITGAV	3.69
P27797	Calreticulin	CALR	CALR	3.35
P01137	Transforming growth factor beta-1	TGF-B-1	TGFB1	3.34
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	3.33
Q9Y6Q6	Tumor necrosis factor receptor superfamily member 11A	TNFRSF11A	TNFRSF11A	2.98
O00300	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	TNFRSF11B	1.73
Q99807	5-demethoxyubiquinone hydroxylase, mitochondrial	DMQH	COQ7	1.36
Q96CX2	BTB/POZ domain-containing protein KCTD12	KCTD12	KCTD12	1.31
P02741	C-reactive protein	CRP	CRP	1.14
P02765	Alpha-2-HS-glycoprotein	AHSG	AHSG	1.08

Supplemental Table 20 – Plasma-“All” subset’s potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
P22301	Interleukin-10	IL-10	IL10	4.43
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	2.37
-	Brain Natriuretic Peptide	BNP	NPPB	2.31
-	Atrial Natriuretic Peptide	ANP	NPPA	1.89
P36222	Chitinase-3-like protein 1	CHI3L1	CHI3L1	1.86
-	N-terminal pro-Atrial Natriuretic Peptide	NTproANP	NPPA	1.39
P13726	Tissue factor	TF	F3	1.37
-	Beta-thromboglobulin	B-TG	PPBP	1.10
-	D-dimer	DD	-	1.08
P05362	Intercellular adhesion molecule 1	ICAM1	ICAM1	1.06
P43235	Cathepsin K	CTSK	CTSK	1.02

Supplemental Table 21 – Serum-“All” subset’s potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
-	Atrial Natriuretic Peptide	ANP	NPPA	1.90
-	Brain Natriuretic Peptide	BNP	NPPB	1.78
-	Relaxin	RLX	RLN2	1.16
P14780	Matrix metalloproteinase-9	MMP-9	MMP9	1.16
Q14116	Interleukin-18	IL-18	IL18	1.03

Supplemental Table 23 – Atrial Appendages-paroxysmal AF subset’s potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	2.80
Q9Y6Q6	Tumor necrosis factor receptor superfamily member 11A	TNFRSF11A	TNFRSF11A	2.46
P06756	Integrin alpha-V	ITGAV	ITGAV	2.38
P01137	Transforming growth factor beta-1	TGF-B-1	TGFB1	1.74
P27797	Calreticulin	CALR	CALR	1.46
O00300	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	TNFRSF11B	1.41

Supplemental Table 22 – Plasma- paroxysmal AF subset’s potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
-	Brain Natriuretic Peptide	BNP	NPPB	2.35
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	1.98
-	Atrial Natriuretic Peptide	ANP	NPPA	1.86
-	N-terminal pro-Atrial Natriuretic Peptide	NTproANP	NPPA	1.03

Supplemental Table 24 – Atrial Appendages-persistent AF subset's potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
P27797	Calreticulin	CALR	CALR	4.22
P01137	Transforming growth factor beta-1	TGF-B-1	TGFB1	3.98
P06756	Integrin alpha-V	ITGAV	ITGAV	3.81
P62937	Peptidyl-prolyl cis-trans isomerase A	PPIaseA	PPIA	2.83
O14788	Tumor necrosis factor ligand superfamily member 11	TNFSF11	TNFSF11	2.35
P49748	Very long-chain specific acyl-CoA dehydrogenase, mitochondrial	VLCAD	ACADVL	2.26
Q9Y6Q6	Tumor necrosis factor receptor superfamily member 11A	TNFRSF11A	TNFRSF11A	2.22
P00441	Superoxide dismutase Cu-Zn	SOD1	SOD1	1.79
P08590	Myosin light chain 3	MYL3	MYL3	1.64
P12821	Angiotensin-converting enzyme	ACE	ACE	1.51
Q15796	Mothers against decapentaplegic homolog 2	MADH2	SMAD2	1.50
O00300	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	TNFRSF11B	1.49
Q06830	Peroxiredoxin-1	PRDX1	PRDX1	1.44
Q9P0J0	NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 13	NDUFA13	NDUFA13	1.33
P30086	Phosphatidylethanolamine-binding protein 1	PEBP-1	PEBP1	1.29
P23528	Cofilin-1	CFL1	CFL1	1.27
P35609	Alpha-actinin-2	ACTN2	ACTN2	1.22
P19429	Troponin I, cardiac muscle	TnIc	TNNI3	1.11
P45880	Voltage-dependent anion-selective channel protein 2	VDAC-2	VDAC2	1.01

Supplemental Table 25 – Plasma-persistent AF subset's potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
-	Brain Natriuretic Peptide	BNP	NPPB	3.44
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	3.05
P15692	Vascular endothelial growth factor A	VEGF-A	VEGFA	2.66
-	Atrial Natriuretic Peptide	ANP	NPPA	1.97
P13726	Tissue factor	TF	F3	1.46

Supplemental Table 26 – Serum-persistent AF subset's potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
P26022	Pentraxin-related protein PTX3	PTX3	PTX3	2.00
-	Atrial Natriuretic Peptide	ANP	NPPA	1.80
-	Brain Natriuretic Peptide	BNP	NPPB	1.50
P25445	Tumor necrosis factor receptor superfamily member 6	TNFRSF6	FAS	1.38
P14210	Hepatocyte growth factor	HGF	HGF	1.34
P01137	Transforming growth factor beta-1	TGF-B-1	TGFB1	1.20

Supplemental Table 27 – Atrial appendages-permanent AF subset's potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
P11021	Endoplasmic reticulum chaperone BiP	HSPA5	HSPA5	4.04
P11177	Pyruvate dehydrogenase E1 component subunit beta, mitochondrial	PDHE1-B	PDHB	2.87
P06753	Tropomyosin alpha-3 chain	TPM3	TPM3	2.04
P10809	60 kDa heat shock protein, mitochondrial	Hsp60	HSPD1	1.92
P12829	Myosin light chain 4	MYL4	MYL4	1.89
P68032	Actin, alpha cardiac muscle 1	ACTC1	ACTC1	1.58
P07951	Tropomyosin beta chain	TMSB	TPM2	1.53
P30084	Enoyl-CoA hydratase, mitochondrial	ECoAh	ECHS1	1.35
P06732	Creatine kinase M-type	M-CK	CKM	1.30
P04792	Heat shock protein beta-1	HspB1	HSPB1	1.01

Supplemental Table 28 – Plasma-postoperative AF recurrence subset's potential biomarker proteins after the bootstrap approach with $p = 50\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
P22301	Interleukin-10	IL-10	IL10	4.42
-	pro-Brain Natriuretic Peptide	proBNP	NPPB	2.63
P15692	Vascular endothelial growth factor A	VEGF-A	VEGFA	2.52
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	2.39
-	Brain Natriuretic Peptide	BNP	NPPB	2.34
-	Atrial Natriuretic Peptide	ANP	NPPA	1.93
P36222	Chitinase-3-like protein 1	CHI3L1	CHI3L1	1.93
-	N-terminal pro-Atrial Natriuretic Peptide	NTproANP	NPPA	1.38
P13726	Tissue factor	TF	F3	1.36
P43235	Cathepsin K	CTSK	CTSK	1.13
-	D-dimer	DD	-	1.10
P05362	Intercellular adhesion molecule 1	ICAM1	ICAM1	1.09
-	Beta-thromboglobulin	B-TG	PPBP	1.08

Supplemental Table 29 – Whole blood-persistent AF subset's potential biomarker proteins after the bootstrap approach with $p = 75\%$.

UNIPROT Code	Full Name	Abbreviation	Gene	Mean Score
P02741	C-reactive protein	CRP	CRP	2.30
-	N-terminal pro-Brain Natriuretic Peptide	NTproBNP	NPPB	1.46
P16109	P-selectin	SELP	SELP	1.03
-	Mid-region pro-Atrial Natriuretic Peptide	MRproANP	NPPA	1.02