Study Title

Detection of Coronary Artery Calcifications by whole blood Transcriptome analyzed by Artificial InTelligence algorithms. (CAC-TRAIT study)

The information contained herein is **strictly confidential** and must not be disclosed, copied, presented for publication or used for any purpose without the prior written authorization of the participating researchers.

Research area: Subclinical atherosclerosis and prevention of cardiovascular events

Clinical phase: Analytical, observational

Protocol version: October 12, 2022

Centers: ENERI Medical Institute, La Sagrada Familia Clinic

CABA, Argentina

Principal Investigator of the study

Santiago Miriuka

Co-principal investigators of the study

Rosana Poggio

Carlos Luzzani

Gastón Rodríguez-Granillo

Bioinformatic analysis

Alejandro La Greca

Coordinator of the on-site study

Bibiana Rubilar

Index

Index 1

1. Contacts 3

2. Abbreviations 3

3. Abstract 4

4. Background and Rationale 6

4.1. Cardiovascular disease 6

4.2. Transcriptomics 7

4.3. Polygenic risk score and its functional derivations 8

4.4. Evidence of alterations in peripheral blood in patients with atherosclerosis: 9

4.5. Artificial intelligence 11

4.6. Use of artificial intelligence in medicine 12

5. Preliminary observations 13

5.1. In silico developments 13

5.2. Preliminary studies 17

6. Study Objectives 18

Primary Objective 18

7. Methods 19

7.1. Study Design 19

7.2. Study Population 20

7.3. Inclusion Criteria 20

7.4. Exclusion criteria. 20

8. Procedures 21

8.1. Elegibility and informed consent 21

8.2. Collection of clinical data and procedures in the baseline stage 21

Peripheral blood sample 22

8.3. Sample processing and sequencing 23

8.4. Bioinformatic and computational analysis of the transcriptomics sample 24

8.5. Analysis of cardiac CT 24

8.7. Patient follow-up **26**
8.7.1. Events of interest in the study during follow-up 26
8.7.2. Verification and adjudication of events during follow-up 27
9. Guarantee and quality control of the data **27**
9.1. Confidentiality of the data 27
9.2. Data quality assurance and control Data 27
9.3. Calculation of the sample size **29**
9.4. Analysis plan **29**
9.5. Control of good clinical practices **30**
9.6. Protocol Deviations **30**
9.7. Publication policy **31**
9.8. Intellectual Property **31**
9.9.1. Direction of the Study **31**
9.10. Financing **32**
10. References **32**

1. Contacts

- Santiago Miriuka (Principal Investigator)

  MultiplAI Health

  Telephone: 1161615422

  e-mail: santiago@multiplaihealth.com
- Carlos Luzzani (Co-Investigator)

  MultiplAI Health

  e-mail: charlie@multiplaihealth.com
- Rosana Poggio (Co-Investigator)

  e-mail: rosana.poggio@multiplaihealth.com
- Gastón Rodríguez-Granillo (Co-Investigator)

  e-mail: grodriguezgranillo@gmail.com

2. Abbreviations

ABC: Area Under the Curve

CAC: Coronary artery calcification

PBMC: Peripheral blood mononuclear cells

DL: Deep learning

CVD: Cardiovascular disease

AI: Artificial intelligence

RNA -Seq: Massive RNA sequencing

CT: Computed tomography

3. Abstract

At present, it is possible to obtain a large amount of information from a biological sample employing technologies known collectively as *omics*. One of them in particular, transcriptomics, can measure all the RNA molecules in the cells or tissues that are the object of analysis. The set of these RNA molecules is known as the transcriptome and represents the functional expression of the DNA of said cells or tissues. However, one of the problems that this large mass of data presents is its interpretation: the transcriptome is made up of tens of thousands of genes that can vary their expression and conformation and, therefore, their understanding, visualization, and translation of these changes to the medicine are complex. Until now, the way to interpret this molecular expression system was with the principles of classical statistics: those RNA molecules that were differentially expressed in the tissues were the ones that were relevant and were potential positive markers of the differences between the cells or tissues analyzed. While this is correct, it has several problems. First, false positives are frequent since the number of comparisons made is very high. Secondly, these data are thousands when samples are compared, so it is difficult for the doctor to interpret them as a whole since they cannot assess each marker individually. Finally, minor but consistent (and possibly the most frequent) variations of a large group of genes are not observed and not considered at the time of interpretation.

In recent years, artificial intelligence (AI) has developed methods to help us interpret *omics*. The analysis methods collectively known as *machine learning* or *deep learning* (DL), make it possible to consider the salient characteristics (*features*) and their constant variation throughout a population. In this way, algorithms can learn to classify new samples by recognising these characteristics (*features)* and evaluating them as a whole. These classification methods allow one transcriptomics to be distinguished from another one with a high degree of accuracy.

Previous evidence suggests changes in peripheral blood transcriptomics due to alterations predisposing to coronary heart disease. These changes are due to two main reasons. First, exposure to coronary risk factors (known and unknown) determines

changes in gene expression, particularly those related to metabolism, immunity, and platelet aggregation. Second, the development of coronary heart disease (and other complex diseases) is facilitated by predisposition represented by the inheritance of a significant number of variants in the genome. Thus, the trait of a complex disease can be predicted by measuring the polygenic risk score (PRS). Finally, recent studies confirm the association of heredity with exposure, thus combining the two dimensions in a single risk score.

In our preliminary tests, we worked on fine-tuning a measurement in a peripheral blood sample that allows the integration of all the genetics, including heredity as expression. We developed a novel way to analyze a biological sample's transcriptome through convolutional neural networks (*deep learning*). Our findings suggest that our analysis can classify the samples with very high precision. We applied this technology in a preliminary study, in which we found associations between surrogate points for cardiovascular disease and gene expression in peripheral blood. For example, we conducted a clinical study measuring coronary artery calcification (CAC), a clear surrogate marker of atherosclerotic disease and cardiovascular events, closely related to plaque load, of long-established prognostic value, and amenable to accurate noninvasive assessment by computed tomography (CT). In this study, we found clear indications of the clinical usefulness of our methodology.

The present clinical study proposes to identify atherosclerotic transcriptomic patterns from a whole blood sample. Likewise, as a secondary analysis, we will explore the ability of the algorithm to determine the presence of aortic disease and pro-inflammatory cardiometabolic alterations such as hepatic steatosis and surrogate markers of epicardial fat inflammation.

4. Background and Rationale

4.1. Cardiovascular disease

Cardiovascular disease (CVD) is the leading cause of death and disability globally. In 2019, CVD was responsible for approximately 18,600,000 deaths (32.8% of the total), and 80% occurred in low- and middle-income countries. (Mensah et al., 2019). It is possible that after the pandemic, these numbers have increased. The early detection of individuals at high risk of developing cardiovascular disease without symptoms has been and continues to be a challenge for science and health authorities worldwide. Implementing efficient and accessible methodologies for the detection of cardiovascular disease can lead to a radical change in the management of the disease, as well as a significant reduction in health spending associated with it and an extension in the population's life expectancy. The traditional approach to identifying individuals at increased risk of developing CVD or death from this cause is based mainly on the identification of single variables (diabetes, cholesterol, smoking), a group of variables ("Framingham Risk Score") or biomarkers (Hs-CRP, BNP, etc.) that are associated with a higher risk of developing this disease. Despite these advances, these approaches have a limited prediction capacity (Area under the curve 0.60-0.65) (Y. Wang et al., 2019; Rosenberg, 2010). One possible cause of this limited effectiveness is that numerous processes (some known and many still unknown) are involved in developing chronic diseases that cannot be fully captured by analysing one or a few predictor variables. On the other hand, cardiovascular imaging is helpful for the early detection of CVD development (for example, carotid ultrasound and detection of CAC by CT). However, it requires expensive equipment and specialised personnel, so its use in clinical practice is scarce, and its implementation as a population screening tool is unlikely.

In recent years, the detection of genetic variants predisposing to coronary heart disease development has made significant progress. Like many other complex diseases, coronary heart disease has a hereditary component in its development. Classically, the

hereditary component was easily identifiable in two ways. Some patients develop symptoms at an early age (for example, a myocardial infarction under 45-50 years). Other patients have evident phenotypic traits (eg, familial hypercholesterolemia) that predispose them to cardiovascular events. These latter cases primarily respond to specific variants in the genome (monogenic diseases), but they are a minor part of the spectrum of patients with cardiovascular disease. Many patients with a genetic predisposition to cardiovascular disease inherit an extensive set of genomic variants. Although this polygenic inheritance is recognized, many genetic variants found concerning a phenotype do not have a clear causal relationship with the associated pathology. They are only associated with the phenotype because they are part of a haplotype and thus relate, either in *cis* or *trans*, to other causal variants. Coronary heart disease associated with specific haplotypes is becoming an effective method for cardiovascular prediction.

4.2. Transcriptomics

Cell DNA ensures that the biological information of a species is transmitted through generations. However, the interaction of DNA with the environment is established in different ways. The most direct and recognised is through the conversion of its code through the expression of RNA, a process known as transcription. The set of all expressed transcripts at a given time is known as the transcriptome. Therefore, it follows from these concepts that transcriptomics is the intersection of that information reserve that is DNA with modulation by signals from the ecosystem.

Within the structure of DNA are units known as *genes*. Until a few years ago, a gene was considered a DNA sequence that is transcribed into RNA and then translated into a protein. With this definition, it was found that the human genome contains approximately 22,000 genes, not very different from the number of genes that, for example, a fly contains (around 14,000) if we consider the phenotypic differences between one species and another. However, the definition of *gene* and the number have changed recently, possibly reaching as high as 100,000 (Salzberg, 2018). There are many sequences in DNA that are transcribed but not translated, known globally as

non-coding-RNA. These non-coding genes are the ones that mostly allow phenotypic differences between species, individuals, tissues and cells. Finally, the genes represented in the DNA sequence undergo significant changes after being transcribed, such as *splicing* or circularisation, which expands the possible number of differences between biological samples. In short, RNA represents the dynamic and active part of genetic information, profoundly modulated by the environment.

Technology has evolved to analyse the whole transcriptomics produced in a sample at a given time. This was partially done until a few years ago through a study known as a *microarray*, but it is currently done through massive RNA sequencing (RNA-Seq). In it, all the RNAs present at a given moment in a cell or tissue can be identified. However, the amount of information produced is enormous and must be ordered. For this, different *software* allows alignment with the human genome to identify each of the genes and count them to determine their expression level.

The development of massive sequencing and transcriptomics must also be seen in another context: lower costs. A few years ago, the cost per sample was well over a thousand dollars. However, the market launch of equipment with superior capabilities has markedly lowered the costs per sample. Therefore, this is a factor of paramount importance for developing clinical studies with tens or hundreds of patients. The next five years will see a continued expansion in reports of using RNA-Seq in patients. The determination of its clinical usefulness will depend on adequate analysis, for which the development of techniques based on artificial intelligence will facilitate its interpretation.

4.3. Polygenic risk score and its functional derivations

As previously mentioned, heredity plays a relevant role in some patients with coronary disease. In 2007, Samani et al. described the association of three haplotypes with early coronary disease (Samani et al.). Since then, multiple studies have shown the existence of specific haplotypes for different chronic diseases, including atherosclerosis and coronary heart disease (Musunuru and Kathiresan). These haplotypes have dozens of

genomic variants associated with a phenotype, each slightly associated with it. This minimal association is possibly the product of the absence of a direct causal relationship, and the existence of the minimal relationship is due to the absence of independent segregation of the alleles (*linkage disequilibrium*).

Based on the knowledge of the minimum association between variants and the possibility of massive DNA sequencing at low cost, population studies were carried out that allowed the development of risk scores for the development of complex diseases based on the sum of the minimum inherited risks. Thus, in recent years *scores* known as PRS (*Polygenic Risk Score*) have been designed, which result, for example, from the algebraic sum of the odds ratios of each of the variants. The higher the score, the higher the risk. This way of adding minimal risks proved efficient in discriminating against individual risks. Thus, a person in the highest risk decile has the same risk as a patient with familial hypercholesterolemia (Inouye et al., 2018; Klarin & Natarajan, 2021).

The clinical utility of the PRS is in the phase of clinical application, and it may be routine in the coming years. However, the PRS only measures the hereditary component of risk. Different researchers have proposed using PRS modified by including a component related to gene expression to establish a risk that also involves the environmental component of the risk, which may come from the association of genetic variants with epigenetic or transcriptome variations (Cano-Gamez & Trynka, 2020). In this way, the concept of *Transcriptome-wide Association Studies* (TWAS) was born, which measures risk based on the combination of inherited genetic variants and the expression of genes close to the same variants (Wainberg et al., 2017; Wainberg et al., 2019; Bhattacharya et al., 2020).

4.4. Evidence of alterations in peripheral blood in patients with atherosclerosis:

Atherosclerosis is known as a vascular disease caused by the deposition of cholesterol in inflammatory cells in the middle layer of the arteries. The consequences of

these deposits are well known: exposure to lipid-containing inflammatory cells after plaque rupture produces most of the acute intravascular thrombotic events, including myocardial infarction, some types of stroke, and peripheral arterial disease. (Wolf & Ley, 2019; Geovanini & Libby, 2018). Non-occlusive events eventually lead to low blood flow ischemia.

From the preceding, the necessary interaction between inflammatory cells, metabolism and the vascular wall in the pathophysiology of atherosclerotic disease is deduced. This complex interaction is often measured in clinical practice by surrogate biomarkers of vascular status and the odds of unfavourable outcomes of atherosclerotic disease. These markers can be representative of metabolic changes (cholesterol, blood glucose, free fatty acids, etc.), inflammatory (C-reactive protein, interleukins, leptin, etc.) or coagulation (fibrinogen, PAI-1, etc.). Furthermore, many still unknown markers are present in peripheral blood, originating from new cardiovascular risk factors described in recent years, such as the intestinal microbiome (Barrington & Lusis, 2017; Komaroff, 2018) or air pollution (Jilani et al., 2020). In addition, these markers can be of different molecular types: protein, RNAs of various shapes, or even small molecules.

Among the possible biomarkers mentioned in the previous paragraph are RNA molecules. Blood cells are subjected to all kinds of signals from the tissues they perfuse. These signals impact circulating cells, or even progenitors in the bone marrow, thereby modifying gene expression. Thus, the different states of a disease can present a set of differentially expressed genes. Various studies have confirmed the existence of genes expressed in patients with atherosclerotic disease. Alterations in gene expression in blood cells are evidenced in different states of predisposition to atherosclerosis (Ding et al., 2015), prognostic evolution (Holvoet al., 2016), or response to treatment with statins (Kini et al., 2017). Furthermore, the expression of a set of genes has been associated with clinical variables to predict the severity of atherosclerotic disease (Y. Wang et al., 2019). Therefore, these studies show the existence of alterations in gene expression in patients with cardiovascular disease. It is currently claimed that these changes can be used in the

future as markers of the disease and its evolution (Marín de Evsikova et al., 2019; Chen & Stewart, 2016; Friede et al., 2015; Rhees & Wingrove, 2015).

4.5. Artificial intelligence

Artificial intelligence (AI) has spread widely in recent years thanks to the development of new algorithms that allow enormous growth in classification capabilities. The possible applications of this in many areas of human activity have led to AI being considered a new milestone in the evolution of humanity. In particular, for medicine, it can mean a stage of enormous changes in how it is practised (Topol, 2019).

AI can be defined as a set of algorithms and equations that can mimic certain cognitive behaviours of human beings. While the term suggests reasoning capabilities, AI is currently only limited to a human (or even super-human) ability to interpret and classify big data. For this purpose, mathematical algorithms are used, collectively known as *machine learning*. One, in particular, has stood out for some years, known as the neural network. This algorithm involves many parallel and serial linear equations, each modified by one or more nonlinear equations, forming a network of equations, ending in a nonlinear equation that determines a probability. The latter is used for probabilistic classification. Although this algorithm has been known for decades, specific changes in recent years have allowed its enormous expansion, leading them to be known as *deep learning (LeCun et al., 2015)*. These changes are mainly of two types: first, the ability for the network to correct itself was developed in the 1980s by employing an algorithm called *backpropagation (Rumelhart et al., 1986); Drucker & Le Cun, 1992)*. This algorithm allows the neural network to adjust in successive cycles to modifications that progressively improve its predictive capacity (*training*) in successive cycles. The second reason for the expansion of *deep learning* corresponds to the development of *hardware* that makes it possible to carry out a large number of simple calculations (linear regressions) in a short time, mainly through the adaptation of "GPUs" (*Graphical Processing Units*) for use in AI. From these two advances, the science of *deep learning* has expanded remarkably, with increasing participation in processes of daily life.

4.6. Use of artificial intelligence in medicine

Many examples have emerged in recent years about the potential of AI in medicine. One of the most important studies was published in 2017, in which an algorithm was fed photographs of skin lesions, including different forms of skin cancer. The algorithm was able to differentiate those malignant lesions with higher accuracy than a group of dermatologists (Esteva et al., 2017). Another study of equal importance demonstrated a high capacity of DL to distinguish the presence of tumour cells in histopathological images of axillary lymph nodes of patients with breast cancer (Ehteshami Bejnordi et al., 2017). These two examples are a sample of the impact of advances in development in this area: a trained algorithm can automatically distinguish alterations present in an image, regardless of the doctor. The enormous potential for diagnostic help of these algorithms is an impetus for research on them.

Although the use of AI in medicine is in its infancy and the research stage, possible applications of AI in medicine can be grouped into the following spaces:

- Radiology: AI algorithms can be fed with imaging studies, and they can learn characteristics associated with abnormal defects. Based on these capabilities, there is a growing number of potential AI applications in this field, and some of them have already been approved for clinical use. Soon, many of the most common images in radiology will be previously scanned by an algorithm to preliminarily identify if there is an alteration, alleviating the work of the doctor who will only focus on those that contain a pathological alteration. Various studies have shown the usefulness of DL in multiple areas of the cardiovascular field, from a risk stratification and triage tool to the automatic evaluation of ventricular volumes in magnetic resonance imaging and the coronary calcium score by CT.
- Body images: Successful examples in this area are also expected due to DL's good results in image processing. For example, the interpretation of skin lesions is feasible with DL. Another example is the analysis of fundus images, possibly predicting cardiovascular risk. But in addition, neural networks can recognise

whether an eye fundus image is of a man or a woman with an accuracy of approximately 98%, which speaks to their ability to extract imperceptible features from an image. It is not known precisely what characteristics define the differences in the images, but they may be related to the thickness of the arteries or another variation imperceptible to the human eye (Poplin et al., 2018; Nent et al., 2019).

- Continuous signals: This is the case of the use of AI for interpreting data coming from continuous acquisition signals. Examples of them are, for example, the interpretation of an electrocardiogram or electroencephalogram by a DL algorithm (Hannun et al., 2019; Craik et al., 2019).
- Language processing: The interpretation of language is being revolutionised by DL. The automated analysis of medical records or the patient vocabulary are examples of how it can be applied. There are multiple reports where they identify emotions, including depressive states, from the language used by a subject (Liang et al., 2016).

5. Preliminary observations

5.1. In silico developments

In our laboratory, we have developed a methodology for transcriptome analysis based on extensive bioinformatic processing of a biological sample and its subsequent interpretation employing a *deep learning algorithm.* In this way, we can incorporate all the genes read in a sequence and perform a classification using a supervised method. This algorithm has proven robust and highly accurate in our preliminary tests. The development comprises a *pipeline* that takes the RNA-seq sample in its crudest state and analyses it entirely, including coding and non-coding RNAs, circular RNAs and the blood microbiome. The algorithm is trained with all the samples of a cohort, previously classified, and later, the algorithm is validated with 10 or 20% of the samples.

The analysis of the efficiency of the algorithm was performed with different RNA-seq databases available in digital repositories. First, we used the GTEx database, which has taken *postmortem* from dozens of people and performed sequencing of them. In total, 30 different tissues containing 11,866 sequences were analyzed. The analysis of our algorithm achieved a correct prediction in 99.1% of the cases (figure 1). The few errors correspond to potentially similar tissues (for example, breast with fatty tissue).

Subsequently, we analyzed the Genomic Data Commons (GDC) database, which has RNA sequencing of biopsies from different tumors. We analyzed 10,000 biopsies distributed in 34 tumors. Again, the predictive capacity of the algorithm was very high, reaching 97.5% (figure 2). These two analyses show that the algorithm has a great classification capacity between multiple samples.

We then performed a sensitivity analysis of the algorithm. To do this, we started with 100 random samples taken from brain tumors from the GDC database, 50 of them corresponding to GBM and 50 of them to GL. With them, new training was carried out in which we achieved a precision of 100%, which translates into an area under the curve (ROC) of 1. Successively, we removed samples at random, performing effective training with only 30 samples (15 of each group). Therefore, the algorithm has a very efficient behaviour with a low number of clinical cases. This reflects the fundamental differences in gene expression between the two groups of tumors.

Confusion matrix

Actual \ Predicted	Adipose Tissue	Adrenal Gland	Bladder	Blood	Blood Vessel	Brain	Breast	Cervix Uteri	Colon	Esophagus	Fallopian Tube	Heart	Kidney	Liver	Lung	Muscle	Nerve	Ovary	Pancreas	Pituitary	Prostate	Salivary Gland	Skin	Small Intestine	Spleen	Stomach	Testis	Thyroid	Uterus	Vagina
Adipose Tissue	148	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Adrenal Gland	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bladder	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Blood	0	0	0	110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Blood Vessel	1	0	0	0	171	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Brain	0	0	0	0	0	345	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Breast	3	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Cervix Uteri	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Colon	0	0	0	0	0	0	0	0	116	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Esophagus	0	0	0	0	2	0	0	0	0	197	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fallopian Tube	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Heart	0	0	0	0	0	0	0	0	0	0	0	120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kidney	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Liver	0	0	0	1	0	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lung	1	0	0	0	0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Muscle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	117	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nerve	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	84	0	0	0	0	0	0	0	0	0	0	0	0	0
Ovary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	1
Pancreas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0
Pituitary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0
Prostate	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0
Salivary Gland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0
Skin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	228	0	0	0	0	0	0	0
Small Intestine	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0
Spleen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0
Stomach	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	0	0	1
Testis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	53	0	0	0
Thyroid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	88	0	0
Uterus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	16	0
Vagina	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19

Figure 1: Confusion matrix for normal tissue samples. The graph shows the intersection of the algorithm's prediction and the actual sample. Thus, the descending central diagonal from left to right shows the coincidence between the real and the predicted. Outside this diagonal, the numbers represent the algorithm's classification errors. As can be seen, they are infrequent.

The above results were obtained with RNA-Seq samples from solid tissues (biopsies). As mentioned above, the cells in the blood can present modifications in their transcriptomics in the presence of a solid tumor or systemic disease. Therefore, below we perform analyzes of different datasets in different contexts.

First, we analyzed the samples obtained by Best et al., performed on patients with different solid tumors, including breast, lung, colon, and glioblastoma multiforme cancer (Best et al., 2015). The transcriptomics of these patients was performed on platelets. In all the cases analyzed, the AUC for detecting the presence of the tumor was 1.

Confusion matrix

Actual \ Predicted	acute_myeloid_leu_AML	adrenal_ACC	adrenal_PCPG	bladder_BLCA	brain_GBM	brain_LGG	breast_BRCA_ductalandlobular	cervix_CESC_squamous	colon_COAD_adenoCA	esophagus_ESCA_adenoca	esophagus_ESCA_squamous	kidney_KICH_adenoCA	kidney_KIRC_adenoCA	kidney_KIRP_adenoCA	kidney_WT	larynx_HNSC	leumyelo_LAML	liver_CHOL	liver_LIHC	lung_LUAD_adenoCA	lung_LUSC	melanoma_SKCM	ovary_OV	pancreas_PAAD	prostate_PRAD_adenoCA	stomach_STAD	testis_TGCT	thymus_THYM	thyroid_THCA	tongue_HNSC	utero_UCEC_adenoCA	utero_UCEC_cystic
acute_myeloid_leu_AML	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
adrenal_ACC	0	9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
adrenal_PCPG	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bladder_BLCA	0	0	0	36	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
brain_GBM	0	0	0	0	22	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
brain_LGG	0	0	0	0	1	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
breast_BRCA_ductalandlobular	0	1	0	0	0	0	118	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cervix_CESC_squamous	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
colon_COAD_adenoCA	0	0	0	0	0	0	0	0	57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
esophagus_ESCA_adenoca	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
esophagus_ESCA_squamous	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidney_KICH_adenoCA	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidney_KIRC_adenoCA	0	0	0	1	0	0	0	0	0	0	0	0	45	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidney_KIRP_adenoCA	0	0	0	0	0	0	0	0	0	0	0	2	1	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidney_WT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
larynx_HNSC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0	0
leumyelo_LAML	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
liver_CHOL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
liver_LIHC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0
lung_LUAD_adenoCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	3	0	0	0	0	0	0	0	0	0	0	0
lung_LUSC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	50	0	0	0	0	0	0	0	0	0	0	0
melanoma_SKCM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	50	0	0	0	0	0	0	0	0	0	0
ovary_OV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0
pancreas_PAAD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
prostate_PRAD_adenoCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65	0	0	0	0	0	0	0
stomach_STAD	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0
testis_TGCT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
thymus_THYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
thyroid_THCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65	0	0	0
tongue_HNSC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
utero_UCEC_adenoCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	46	4
utero_UCEC_cystic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	12

Figure 2: Confusion matrix for tumor tissue samples. The graph has the same meaning as in the previous figure. Again, the numbers of mistakes outside the diagonal line are very few

We then analyzed the samples from Cassetta and collaborators on monocytes from patients with breast or endometrial cancer, compared to healthy subjects (Cassetta et al., 2019). Once again, the prediction capacity of the algorithm was 100%. Finally, we analyzed the *dataset* generated by Scharer et al. on patients with lupus compared to healthy subjects (Scharer et al., 2019). Again, classification was excellent (AUC of 1).

In summary, the pipeline generated and the analyses performed to demonstrate its ability to classify transcriptomics samples. The algorithm proves to be robust and extremely sensitive, possibly for two reasons. First, the number of variants it considers is enormous: approximately 200,000 different forms of genes can be evaluated. Secondly,

the analysis capacity of neural networks makes it possible to consider this great mass of information with an unprecedented capacity to interpret this data. These findings then lead us to propose the hypothesis that our pipeline may be useful in detecting tumour pathologies in peripheral blood transcriptomics.

5.2. Preliminary studies

We previously developed a clinical protocol with the aim of perfecting the methodology that will be used in this study. The initial goal was to develop blood collection, RNA isolation, library generation for sequencing, and RNA sequencing. Briefly, the preliminary results of these steps, performed on 540 samples from 3 different studies, are described below.

First, a robust methodology was established to obtain an RNA sample from peripheral whole blood (without cell separation). Samples were collected in PaxGene RNA tubes through antecubital vein puncture. No adverse effects were observed in this procedure. Samples were stored at -20°C until RNA extraction. This step was performed through the use of a specific kit (Thermo-Fischer). As a result, it was possible to consistently obtain RNA samples in sufficient quantities for sequencing in all cases. Then, the library was built, a process in which the RNA is transformed into cDNA (complementary DNA), from which the RNA corresponding to ribosomes and globin are extracted, fragmented, and finally the adapters are added at the ends. Again, the library construction was developed in a robust way for all the analyzed samples.

Subsequently, RNA sequencing was performed at a depth of 100 million reads. Between 20 and 25 thousand genes expressed in peripheral blood were detected in all the samples (n=540). In a series of patients, computed tomography was performed to determine the degree of calcification of the coronary arteries (Coronary Calcium Score, CCS). An analysis of gene expression in peripheral blood was performed and genes related to the degree of coronary calcification were found.

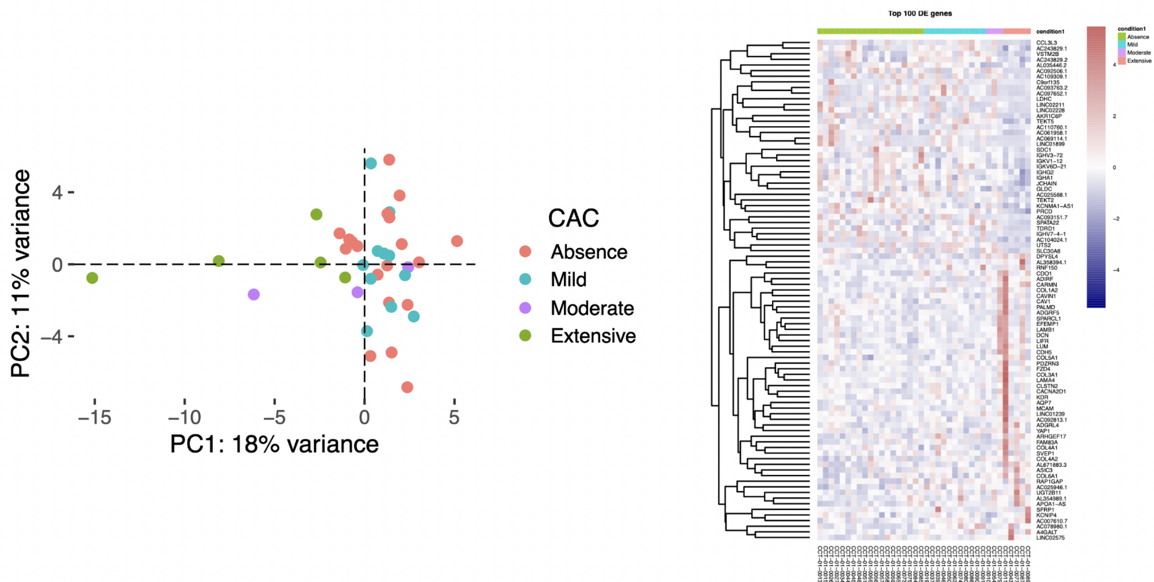

Figure 3: Preliminary results. We analyzed 38 samples from patients who had coronary calcium measured. The left panel shows the result of a PCA analysis where a moderate but significant reduction in dimensionality is found based on the extent of coronary calcium. The heatmap shows the differentially expressed genes between the two groups.

6. Study Objectives

Primary Objective

The primary objective of the study is to evaluate the diagnostic accuracy of peripheral whole blood transcriptomes analyzed by an artificial intelligence algorithm to detect the presence and extent of coronary calcification in individuals without a history of known cardiovascular disease.

Specific Primary Objectives

1. To classify transcriptomes according to the presence and extent of coronary calcification by noncontrast cardiac CT.

2. Train artificial intelligence algorithms with information from the transcriptome to recognize the presence and extent of calcification in coronary arteries.

3. To determine the degree of diagnostic accuracy of the method to detect the presence and extent of coronary calcification non-invasively.

Secondary objectives

1. To determine the degree of diagnostic accuracy of the method to detect the presence and extent of:
    - Aortic
    - calcification Aortic valve calcification
    - Hepatic steatosis
    - Volume of epicardial fat
    - Volume of high-density epicardial fat (upper quartile)
    - Enlargement of cardiac cavities To
2. To explore differences in transcriptomics of peripheral blood between patients who did and did not develop a CV event during follow-up.
3. To explore different electrocardiographic patterns according to the results of cardiac CT and peripheral blood transcriptomics.
4. Generate a polygenic risk score (PRS) from the transcriptome and/or exome sequencing and associate it with the transcriptome and coronary calcification.

7. Methods

7.1. Study Design

This trial will be a prospective observational study. A convenience sample will be carried out to include 800 patients who attend the ENERI Medical Institute, La Sagrada Familia Clinic, to undergo a cardiac CT without contrast (coronary calcium score), either for a medical indication or volunteers to an assessment of cardiovascular risk. The study will have a baseline stage in which a clinical evaluation will be performed, blood samples will be drawn for transcriptome analysis and laboratory analysis. Then, a DNA sample

obtained by swabbing the buccal mucosa will be taken. Subsequently, a non-contrast gated cardiac CT will be performed to assess the presence and extent of coronary calcification and other results of interest. At the end of patient enrollment, biological samples will be sequenced for in silico evaluation of the results. Finally, a 5-year clinical (telephone) follow-up will be carried out to collect data on the incidence of fatal and non-fatal cardiovascular events.

7.2. Study Population

7.3. Inclusion Criteria

- Men and women between 30 and 75 years attending the institution for a non-contrast cardiac CT scan (coronary calcium score) either by medical indication or voluntary for cardiovascular risk assessment.
- Signature of informed consent form

7.4. Exclusion criteria.

- Previously known chronic renal or hepatic insufficiency.
- Active chronic lung disease, defined as: exacerbated asthma, exacerbated COPD, or pulmonary fibrosis.
- Active and/or previous cardiovascular disease, defined as: previous acute myocardial infarction, stable or unstable angina, cerebrovascular accident, history of vascular interventions (coronary or other territory), heart failure, severe cardiomyopathies or valvulopathies.
- Uncontrolled hyper or hypothyroidism.
- Suprarrenal insufficiency.
- Previous surgeries in the last 3 months.

- Severe trauma in the last 6 months, defined as one that involved bone fractures and/or surgical interventions.
- Known active cancer disease or under treatment (acute or preventive), or history of cancer disease without criteria for cure.
- Known autoimmune disease active or in treatment.
- Ongoing pregnancy, postpartum period of less than 12 months or breastfeeding.
- Other serious illness with an estimated life expectancy of less than 12 months (according to the investigators).
- Temperature greater than 37.5°C recorded by a thermometer or any acute infection caused by viruses or bacteria confirmed by a health professional in the previous 30 days.
- Pathologies under immunosuppressive treatment

8. Procedures

8.1. Elegibility and informed consent

Patients will be recruited at the ENERI Medical Institute and at the outpatient clinics of the “La Sagrada Familia” clinic. Patients who meet the eligibility criteria will be approached by the main investigator or one of the study collaborators and the characteristics of the procedure to be performed, the risks of the associated procedures and the scope of the research to be carried out will be explained to them. After providing the necessary information and answering the questions of the participants, they will be formally invited to sign the informed consent.

8.2. Collection of clinical data and procedures in the baseline stage

At the time of inclusion in the study, the following epidemiological, clinical and complementary study data will be collected using questionnaires specially designed for this study:

Clinical data and medical history

- Contact data: Cell phone , landline, phone of an alternative contact, and E-mail address.
- Epidemiological data: date of birth, gender, ethnicity.
- Clinical history: Presence of any of the following pathologies: dyslipidemia, diabetes, high blood pressure, smoking, previous relevant diseases.
- Medications at the time of admission.
- Level of physical activity and diet.

Peripheral blood sample

2.5 milliliters of peripheral blood will be obtained by puncture of the antecubital vein in a PAXgene or Tempus tube (commercial tubes that contain a solution for the preservation of RNA) to obtain RNA. The patient must be fasting for 8 hours. Subsequently, the blood will be kept at room temperature for up to 48 hours, or at 4°C (in a refrigerator) for up to two weeks, until it is transferred to the laboratory, where it will be frozen at -20°C until it is processed (RNA extraction). .

Subsequently, the second additional peripheral blood sample (10 milliliters in EDTA) will be obtained to measure the complete blood count, glycosylated hemoglobin, hepatogram, lipid profile, among others. The primary tube for clinical laboratory analysis will be sent for analysis to a local laboratory (Stambulián).

Oral mucose swab (mouth swab)

To take the sample, the professional will be asked to rub a buccal swab against the inside of the patient's cheeks and gums, and to place it in a dry sterile 1.5ml eppendorf-type tube or with lysis buffer containing it will be labeled with the participant's ID and then place it in the refrigerator at 4°c waiting to be sent to the laboratory, where the DNA extraction will be performed. Subsequently, it will be stored in a freezer at -20°C.

Non-contrast cardiac computed tomography

ECG-triggered non-contrast cardiac CT ("coronary calcium score") will be performed using the same equipment in all patients, a spectral multidetector tomograph (IQon Spectral CT, Philips Healthcare, Best, The Netherlands). Non-contrast cardiac CT, associated with effective radiation doses of between 1.0 and 1.5 mSv (compared to approximately 5-7 mSv in a conventional chest CT), will be acquired using the following parameters: collimation 32 x 0. 625 mm; voltage 120 kV; mAs adjusted according to weight (<60 kg= 30 mAs; 60 to 80 kg= 50 mAs; >80 kg= 80 mAs); rotation time 270 milliseconds; thickness 2.5 mm, with an acquisition time of approximately 0.18 seconds. Iterative reconstruction algorithms (iDose3) will be used for dose reduction.

12-lead electrocardiogram

All participants will undergo a 12-lead electrocardiogram, which involves the use of 10 electrodes for a correct reading of the electrical potential of the heart during a cardiac cycle. To carry it out, an electrode will be placed on each limb and 6 on the chest, generating 6 leads of the limbs that observe the heart from a frontal plane and 6 precordial leads that observe the heart from a horizontal plane. The risks of an EKG are minimal. Mild skin discomfort or irritation may be felt after the electrodes are removed. There is no risk of electric shock.

8.3. Sample processing and sequencing

Blood samples will be processed after extraction using commercial kits for RNA extraction. Subsequently, the *library*, a procedure by which the transformation of the RNA sample into short cDNA chains for sequencing is known. For this procedure, the commercial TruSeq-RiboZero-GlobinFree kit (Illumina) will be used. Subsequent sequencing (without ribosomal RNA and globin) will be performed by an external sequencing service provider. A sequencing with a read depth of at least 100 million reads will be requested.

8.4. Bioinformatic and computational analysis of the transcriptomics sample

The bioinformatic analysis of the sequenced samples will be performed from the .fastq files containing the raw data. These files will be fed to an *pipeline* analysis It comprises a successive series of bioinformatic procedures, including sample quality control, alignment with the human genome and count of *reads* per identified gene. The transcripts to be identified correspond to messenger RNAs, non-coding RNAs, circular RNAs, and RNAs corresponding to the blood microbiome. The expression quantification of the transcripts is then transformed using a series of proprietary routines, to then be transformed into an image (US patent 62/944,063 (pending)). This image represents both the expression of the transcripts as a function of their expression level and the genomic position as well as the RNA subtype. The set of images, which represent the set of gene expression in peripheral blood, is subsequently entered into a neural network to generate training that classifies the samples according to it. All these bioinformatic and artificial intelligence procedures are carried out using *free use programs (“open access software”)* under a routine created by the study researchers, mainly using python for coding, and pyTorch for training the neural network.

8.5. Analysis of cardiac CT

The presence of CAC and its extension will be evaluated using both ordinal variables (number of segments with CAC and number of affected vessels, and qualitatively: absence, mild, moderate, and severe; and the number of coronary segments with calcification) as continuous variables (Agatston units). The presence and extension of calcifications at the level of the aortic valve and the thoracic aorta (Agatston units) will also be evaluated independently. The extent of aortic calcification will be evaluated in sections that include from the bifurcation of the pulmonary artery as the upper limit to the diaphragm as the lower limit.

Quantitative analysis of SCC will be performed using dedicated software (HeartBeat-CS, Philips Healthcare, Best, The Netherlands). Finally, the volume and density of epicardial fat will be calculated by evaluating the volume of tissue with a density between -190 and -30 HU located in the area between the epicardium and the pericardium, the bifurcation of the pulmonary artery as the upper limit, and the diaphragm as lower limit.

Finally, the density of the hepatic parenchyma will be calculated through a region of interest >100 mm2 at the level of the right and splenic lobes. From these measurements, the presence of hepatic steatosis will be defined from a quotient of less than 1.

At least 20% of the participants (randomly selected) will have a double reading of the main outcome of the study (presence or extent of CACs). This reading will be carried out by an expert external to the group of researchers of the study and blind to the first reading. In case of disagreement, consensus will be sought between the two experts. The aim of this process is to enhance the internal validity of the study.

The epicardial fat volume analysis will be performed by experts from the University of Palermo (Italy).

8.6. Storage of biological samples

The remainder of the RNA samples and all of the DNA samples will be stored in a freezer at -20° C to be used in future research. The biological samples will be coded (see Confidentiality of the data) to maintain the anonymity of the participants and will be stored at FLENI Escobar Headquarters. The storage place has the necessary equipment to ensure their proper preservation. The temperature and other storage conditions will be permanently monitored as part of the quality assurance of the biological samples. It also has an emergency plan with trained personnel with defined responsibilities to attend to alarms and take actions to restore the system in the event of a failure. The freezer that

preserves the biological samples and the monitoring systems are connected to a permanent electrical source, with an emergency generator with enough autonomy to ensure its uninterrupted operation.

Since the stored samples will not be used in this study, specific consent will be requested for their storage and subsequent analysis..

8.7. Patient follow-up

Patients included in the study will be contacted annually by phone call to collect the following data:

- Vital status.
- Cardiovascular events: New episodes of hospitalization due to acute ischemic syndrome, cerebrovascular event, heart failure or revascularization.
- New diagnoses of relevant chronic diseases.

If they cannot be contacted, the "alternative means of contact" provided by the patient will be used, in order to know their vital status or obtain a new contact number.

8.7.1. Events of interest in the study during follow-up

- Death from any cause
- Acute ischemic syndrome: Diagnosis of acute myocardial infarction or unstable angina.
- New diagnosis of heart failure with preserved or impaired ventricular function.
- Vascular revascularization: performance of myocardial revascularization surgery, coronary angioplasty or peripheral vascular angioplasty.
- Cerebrovascular event: Diagnosis of cerebrovascular accident (“stroke”) or transient ischemic attack.
- Cancer: new diagnosis of any type of cancer

8.7.2. Verification and adjudication of events during follow-up

When an event of interest is reported during study follow-up, the "phase of verification" that consists of obtaining data on: type of event, date, place where it was treated and its corresponding source document (death certificate, epicrisis, complementary study, evolution of clinical history, etc.) where the diagnosis of the event appears with a doctor's signature. The study event adjudication committee will be made up of three medical investigators (RP, SGM, GG). Each of the members will analyze the documentation sent for the case separately and will adjudicate the event. If there is no agreement, it will be awarded by majority.

9.Guarantee and quality control of the data

9.1. Confidentiality of the data

The data of the patients will be collected anonymously from the moment of signing the informed consent. To do this, a representative code of the patient will be created, which will include the initials of the study, followed by a hyphen and the inclusion number. The clinical documentation of the patient and the tube containing the blood of the patients or control subjects will contain only this information. The informed consent will be kept in a different location from the clinical information and under lock and key.

The transcriptomics data will remain anonymized and will receive the name designated at the time of inclusion in the study. The data will be deposited and protected on outsourced servers in compliance with the European regulations of the General Data Protection Regulation (GDPR). AWS is currently used to safeguard the information.

9.2. Data quality assurance and control

a. Data entry and storage

electronic platform RedCap. This electronic data capture platform is covered by security systems, in compliance with current regulations on data protection. The access control will

be configured in such a way that there are two owners with the project administrator role, who will be in charge of authorizing the access of the study personnel, as necessary, and with access levels according to the roles and responsibilities. The electronic forms that will be used for data upload will be specifically designed for this study. The data will be entered by the researcher at the recruitment center, via the web. The data will be stored on a central outsourced server (AWS).

b. Data quality control monitoring

Data storage and processing will be centralized in xx. The personnel in charge of entering the data will be trained in the use of RedCap for this purpose.

The following quality controls will be carried out:

- Validation rules generated according to the nature of the variables, programmed through the data validation modules available in RedCap.
- Automatic query generation (queries) according to pre-established validation rules, to be answered by researchers on the sites.
- Manual design of explanatory queries, to be answered by the researchers on the sites.
- Registration system, history and query control (queries).
- Periodic reports on the status of data entry, query generation and response, on a weekly basis.

c. Elaboration of the manual of procedures

A manual of procedures will be developed where the procedures for the implementation of the baseline data collection questionnaire, blood extraction and follow-up will be described.

d. Training and certification

In order to participate in the study, all staff must attend a training session prior to the start of fieldwork. Training sessions will include all aspects of the protocol and procedures manual regarding enrollment, follow-up, protocol, and measurement

procedures. Those who have acquired the expected skills will be certified to carry out the study procedures. Staff will be retrained if necessary.

e. Site Monitoring Visits and Audits

Monitoring visits and audits will be conducted at the health centre in order to follow up on the progress of the fieldwork, identify problems and carry out retraining if necessary.

9.3. Calculation of the sample size

The calculation of the number of patients to demonstrate a positive effect for the use of peripheral blood transcriptomics for the diagnosis of the presence of severe coronary calcification is carried out based on previous experiences by the group of researchers. There is still no standardized methodology in the medical literature to perform the sample calculation of this type of analysis, since it is very recent and is in the exploration phase.studies *silico* carried out by this group of researchers and the preliminary clinical studies carried out previously, it was observed that at least a sample of 800 patients is necessary for a robust training of the artificial intelligence algorithm that detects transcriptional differences between patients with or without a solid tumor. (for example, breast tumor). Therefore, 800 patients will be included based on the minimum number necessary to perform an adequate training of an artificial intelligence algorithm.

9.4. Analysis plan

For the description of the general characteristics of the study population, descriptive statistics such as means, medians and proportions will be used. To calculate the diagnostic performance of the method, the following will be calculated:

- True Positives (TP) or the probability that the test is positive since the disease is present

- True Negatives (TN) or the probability that the test is negative since the disease is not present.
- False Positives (FP) or Type I Error
- False Negatives (FN) or Type II Error
- Sensitivity: VP/VP+FN
- Specificity: VN/ VN+FP
- Precision or accuracy: VP/VP+FP or sensitivity + (1-specificity)

The diagnostic precision can vary between 0.5 and 1, where 1 represents a perfect diagnostic value and 0.5 is a test without diagnostic discriminatory capacity. An area under the curve (AUC) will also be performed to determine the diagnostic capacity of the trained algorithm.

9.5. Control of good clinical practices

The proposed trial will be conducted following the recommendations of the International Council for Harmonization (ICH) for clinical studies (E6R2) (*ICH Official Web Site : ICH*, nd). In particular, the control of good clinical practices will be supervised in those elements that concern this type of evaluation study of new technologies, including patient confidentiality and informed consent. In addition, the standards suggested by the same regulations will be met (E18, (*ICH Official Web Site : ICH*, nd)).

All study investigators have been certified in GCP guidelines.

9.6. Protocol Deviations

Investigators in this study will strictly monitor compliance with good clinical practice. Any deviation from the protocol will be reported to the ethics committee and the case will be individually analyzed in order to determine the procedures to follow.

9.7. Publication policy

Once the study is finished, it will be published in different ways. First, it will be released to the scientific and medical community in the form of a conference abstract, in order to rapidly communicate the results of the conference to those interested. It will subsequently be published as a paper in an international peer-reviewed journal.

9.8. Intellectual Property

The methodology to be applied in this study has been submitted to the Patent Office for legal protection (preliminary patent US 62/944,063). The intellectual property of the study remains in the hands of MultiplAI Health.

9.9. Study Organization

9.9.1. Direction of the Study

The principal investigator of the project is Dr. Santiago Miriuka from the FLENI institute. Dr. Rosana Poggio and Dr. Carlos Luzzani will be the co-principal investigators. Dr. Gastón Granillo will be the principal investigator of the center (La Sagrada Familia) who will provide logistical support and will be responsible for the entire implementation of the study on site. A CONICET fellow will be the coordinator of the logistics and processing of the biological samples of RNA and DNA. Post-doctoral fellow Alejandro La Greca will be in charge of bioinformatic analysis. María Olivera will be the general coordinator and Monitor of the study and Bibiana Rubilar the coordinator of the study on site.

Field Coordinating Center

The researchers will be in charge of supervising and controlling that all the procedures are carried out following the recommended standards during the implementation of the study, as well as the follow-up of the patients.

Communications

Continuous communication will be maintained between the group of study researchers by telephone, email and whatsapp.

9.10. Financing

The financing to cover the processes and materials necessary to carry out this study will be provided entirely by MultiplAI Health LTD.

10. References

Barrington, WT, & Lusis, AJ (2017). Association between the gut microbiome and atherosclerosis. *Nature Reviews Cardiology*, *14*(12), 699–700. https://doi.org/10.1038/nrcardio.2017.169

Best, MG, Sol, N., Kooi, I., Tannous, J., Westerman, BA, Rustenburg, F., Schellen, P., Verschueren, H., Post, E., Koster, J., Ylstra, B., Ameziane, N., Dorsman, J., Smit, EF, Verheul, HM, Noske, DP, Reijneveld, JC, Nilsson, RJA, Tannous, BA, … Wurdinger, T. (2015). RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell*, *28*(5), 666–676. https://doi.org/10.1016/j.ccell.2015.09.018

Bhattacharya, A., Li, Y., & Love, MI (2020). *MOSTWAS: Multi-Omic strategies for transcriptome-wide association studies*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/2020.04.17.047225

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00424

Cassetta, L., Fragkogianni, S., Sims, AH, Swierczak, A., Forrester, LM, Zhang, H., Soong, DYH, Cotechini, T., Anur, P., Lin, EY, Fidanza, A., Lopez-Yrigoyen, M., Millar, MR, Urman, A., Ai, Z., Spellman, PT, Hwang, ES, Dixon, JM, Wiechmann, L., … Pollard, JW (2019). Human Tumor-Associated Macrophage and Monocyte Transcriptional Landscapes Reveal Cancer-Specific Reprogramming, Biomarkers, and Therapeutic Targets. *Cancer Cell*, *35*(4), 588-602.e10. https://doi.org/10.1016/j.ccell.2019.02.009

Chen, H.-H., & Stewart, AFR (2016). Transcriptomic signature of atherosclerosis in the peripheral blood: Fact or fiction? *Current Atherosclerosis Reports*, *18*(12).

https://doi.org/10.1007/s11883-016-0634-x

Craik, A., He, Y., & Contreras-Vidal, JL (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, *16*(3), 031001. https://doi.org/10.1088/1741-2552/ab0ab5

Ding, J., Reynolds, LM, Zeller, T., Müller, C., Lohman, K., Nicklas, BJ, Kritchevsky, SB, Huang, Z., de la Fuente, A., Soranzo, N., Settlage, RE, Chuang, C.-C., Howard, T., Xu, N., Goodarzi, MO, Chen, Y.-DI, Rotter, JI, Siscovick, DS, Parks, JS, … Liu, Y. (2015). Alterations of a Cellular Cholesterol Metabolism Network Are a Molecular Feature of Obesity-Related Type 2 Diabetes and Cardiovascular Disease. *Diabetes*, *64*(10), 3464–3474. https://doi.org/10.2337/db14-1314

Drucker, H., & Le Cun, Y. (1992). Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, *3*(6), 991–997. https://doi.org/10.1109/72.165600

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, JAWM, Hermsen, M., Manson, QF, Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, MC, Bult, P., Beca, F., Beck, AH, Wang, D., Khosla, A., Gargeya, R., … Venâncio, R. (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, *318*(22), 2199. https://doi.org/10.1001/jama.2017.14585

Esteva, A., Kuprel, B., Novoa, RA, Ko, J., Swetter, SM, Blau, HM, & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Friede, KA, Ginsburg, GS, & Voora, D. (2015). Gene Expression Signatures and the Spectrum of Coronary Artery Disease. *Journal of Cardiovascular Translational Research*, *8*(6), 339–352. https://doi.org/10.1007/s12265-015-9640-6

Geovanini, GR, & Libby, P. (2018). Atherosclerosis and inflammation: overview and updates. *Clinical Science*, *132*(12), 1243–1252. https://doi.org/10.1042/cs20180306

Hannun, AY, Rajpurkar, P., Haghpanahi, M., Tison, GH, Bourn, C., Turakhia, MP, & Ng, AY (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, *25*(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

Holvoet, P., Vanhaverbeke, M., Bloch, K., Baatsen, P., Sinnaeve, P., & Janssens, S. (2016).

Low MT-CO1 in Monocytes and Microvesicles Is Associated With Outcome in Patients With Coronary Artery Disease. *Journal of the American Heart Association*, *5*(12). https://doi.org/10.1161/jaha.116.004207

ICH Official web site : ICH. (nd). Retrieved June 26, 2022, from https://www.ich.org/page/efficacy-guidelines

Inouye, M., Abraham, G., Nelson, CP, Wood, AM, Sweeting, MJ, Dudbridge, F., Lai, FY, Kaptoge, S., Brozynska, M., Wang, T., Ye, S., Webb, TR, Rutter, MK, Tzoulaki, I., Patel, RS, Loos, RJF, Keavney, B., Hemingway, H., Thompson, J., … Samani, NJ (2018). Genomic risk prediction of coronary artery disease in 480,000 adults. *Journal of the American College of Cardiology*, *72*(16), 1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079

Jilani, MH, Simon-Friedt, B., Yahya, T., Khan, AY, Hassan, SZ, Kash, B., Blankstein, R., Blaha, MJ, Virani, SS, Rajagopalan, S., Cainzos-Achirica, M., & Nasir, K. (2020). Associations between particulate matter air pollution, presence and progression of subclinical coronary and carotid atherosclerosis: A systematic review. *Atherosclerosis*, *306*, 22–32. https://doi.org/10.1016/j.atherosclerosis.2020.06.018

Kini, AS, Vengrenyuk, Y., Shameer, K., Maehara, A., Purushothaman, M., Yoshimura, T., Matsumura, M., Aquino, M., Haider, N., Johnson, KW, Readhead, B., Kidd, BA, Feig, JE, Krishnan, P., Sweeny, J., Milind, M., Moreno, P., Mehran, R., Kovacic, JC, … Sharma, S. (2017). Intracoronary Imaging, Cholesterol Efflux, and Transcriptomes After Intensive Statin Treatment. *Journal of the American College of Cardiology*, *69*(6), 628–640. https://doi.org/10.1016/j.jacc.2016.10.029

Klarin, D., & Natarajan, P. (2021). Clinical utility of polygenic risk scores for coronary artery disease. *Nature Reviews Cardiology*, *19*(5), 291–301. https://doi.org/10.1038/s41569-021-00638-w

Komaroff, AL (2018). The Microbiome and Risk for Atherosclerosis. *JAMA*, *319*(23), 2381. https://doi.org/10.1001/jama.2018.5240

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Liang, C., Li, Y., & Luo, J. (2016). A Novel Method to Detect Functional microRNA Regulatory Modules by Bicliques Merging. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(3), 549–556. https://doi.org/10.1109/tcbb.2015.2462370

Marín de Evsikova, C., Raplee, ID, Lockhart, J., Jaimes, G., & Evsikov, AV (2019). The Transcriptomic Toolbox: Resources for Interpreting Large Gene Expression Data within a Precision Medicine Context for Metabolic Disease Atherosclerosis. *Journal of Personalized Medicine*, *9*(2), 21. https://doi.org/10.3390/jpm9020021

Mensah, GA, Roth, GA, & Fuster, V. (2019). The global burden of cardiovascular diseases and risk factors. *Journal of the American College of Cardiology*, *74*(20), 2529–2532. https://doi.org/10.1016/j.jacc.2019.10.009

Musunuru, K., & Kathiresan, S. (2019). Genetics of Common, Complex Coronary Artery Disease. *Cell*, *177*(1), 132–145. https://doi.org/10.1016/j.cell.2019.02.015

Nent, E., Nozaki, C., Schmöle, A.-C., Otte, D., & Zimmer, A. (2019). CB2 receptor deletion on myeloid cells enhanced mechanical allodynia in a mouse model of neuropathic pain. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43858-4

Poplin, R., Varadarajan, AV, Blumer, K., Liu, Y., McConnell, MV, Corrado, GS, Peng, L., & Webster, DR (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, *2*(3), 158–164. https://doi.org/10.1038/s41551-018-0195-0

Rhees, B., & Wingrove, JA (2015). Developing Peripheral Blood Gene Expression-Based Diagnostic Tests for Coronary Artery Disease: a Review. *Journal of Cardiovascular Translational Research*, *8*(6), 372–380. https://doi.org/10.1007/s12265-015-9641-5

Rosenberg, S. (2010). Multicenter Validation of the Diagnostic Accuracy of a Blood-Based Gene Expression Test for Assessing Obstructive Coronary Artery Disease in Nondiabetic Patients. *Annals of Internal Medicine*, *153*(7), 425. https://doi.org/10.7326/0003-4819-153-7-201010050-00005

Rumelhart, DE, Hinton, GE, & Williams, RJ (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Salzberg, SL (2018). Open questions: How many genes do we have? *BMC Biology*, *16*(1). https://doi.org/10.1186/s12915-018-0564-x

Samani, NJ, Erdmann, J., Hall, AS, Hengstenberg, C., Mangino, M., Mayer, B., Dixon, RJ, Meitinger, T., Braund, P., Wichmann, H.-E., Barrett, JH, König, IR, Stevens, SE, Szymczak, S., Tregouet, D.-A., Iles, MM, Pahlke, F., Pollard, H., Lieb, W., … Schunkert, H. (2007). Genomewide Association Analysis of Coronary Artery

Disease. *New England Journal of Medicine*, *357*(5), 443–453. https://doi.org/10.1056/nejmoa072366

Scharer, CD, Blalock, EL, Mi, T., Barwick, BG, Jenks, SA, Deguchi, T., Cashman, KS, Neary, BE, Patterson, DG, Hicks, SL, Khosroshahi, A., Eun-Hyung Lee, F., Wei, C., Sanz, I., & Boss, JM (2019). Epigenetic programming underpins B cell dysfunction in human SLE. *Nature Immunology*, *20*(8), 1071–1082. https://doi.org/10.1038/s41590-019-0419-9

Topol, EJ (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, AN, Knowles, DA, Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, JL, Kyung Im, H., Pasaniuc, B., Rivas, MA, & Kundaje, A. (2017). *Transcriptome-wide association studies: Opportunities and challenges*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/206961

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, AN, Knowles, DA, Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, JLM, Im, HK, Pasaniuc, B., Rivas, MA, & Kundaje, A. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, *51*(4), 592–599. https://doi.org/10.1038/s41588-019-0385-z

Wang, X., Cui, N., Liu, X., & Ming, L. (2019). Identification of a blood-based 12-gene signature that predicts the severity of coronary artery stenosis: An integrative approach based on gene network construction, Support Vector Machine algorithm, and multi-cohort validation. *Atherosclerosis*, *291*, 34–43. https://doi.org/10.1016/j.atherosclerosis.2019.10.001

Wang, Y., Lv, Q., Wu, H., Chen, K., Hong, X., Gu, C., Fu, G., & Zhang, W. (2019). Comparison of MESA of and Framingham risk scores in the prediction of coronary artery disease severity. *Herz*, *45*(S1), 139–144. https://doi.org/10.1007/s00059-019-4838-z

Wolf, D., & Ley, K. (2019). Immunity and Inflammation in Atherosclerosis. *Circulation Research*, *124*(2), 315–327. https://doi.org/10.1161/circresaha.118.313591

Zhang, D., Liu, S., Li, Z., & Wang, R. (2022). Global, regional and national burden of gastroesophageal reflux disease, 1990–2019: update from the GBD 2019 study. *Annals of Medicine*, *54*(1), 1372–1384.

https://doi.org/10.1080/07853890.2022.2074535