Study Title

Whole Blood Transcriptomic Signal According to Coronary Atherosclerotic Plaque Burden Assessed by CT Angiography: CORPLAQ-TRAIT Pilot 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:	July 10th,, 2023
Centers:	1. Instituto Médico ENERI, Clinica La Sagrada Familia, Ciudad Autónoma de Buenos Aires, Argentina 2. Diagnóstico MEDITER, Sanatorio Julio Méndez, Ciudad Autónoma de Buenos Aires, Argentina

Principal Investigator of the study

Rosana Poggio

Co-principal investigators of the study

Santiago Miriuka

Gastón Rodriguez Granillo

Site Principal Investigators

Clínica La Sagrada Familia / Instituto Médico ENERI: Gastón Rodríguez Granillo

Diagnóstico MEDITER, Sanatorio Julio Méndez: Guillermo Ganum

Index

1. Contacts

- Rosana Poggio (Principal-Investigator)

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

  e-mail: grodriguezgranillo@gmail.com
- Santiago Miriuka (Co-Investigator)

  MultiplAI Health Telephone: 1161615422

  e-mail: santiago@multiplaihealth.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

Currently, a wealth of information can be obtained from a biological sample using a group of technologies collectively referred to as omics. Transcriptomics, one of these technologies, measures all RNA molecules in the cells or tissues being analyzed. The collection of these RNA molecules, known as the transcriptome, represents the functional expression of the DNA within the cells or tissues. However, interpreting this vast amount of data poses challenges due to the complexity of the transcriptome, which comprises tens of thousands of genes with varying expression and conformation. Traditionally, this molecular expression system has been interpreted using classical statistical principles, focusing on differentially expressed RNA molecules as relevant markers for distinguishing between analyzed cells or tissues. While this approach has merits, it suffers from several issues, including a high rate of false positives, difficulties in comprehensively interpreting the data, and overlooking minor yet consistent variations in a large group of genes.

In recent years, artificial intelligence (AI) has facilitated the interpretation of omics data through machine learning and deep learning (DL) techniques. These methods enable the identification of salient features and their constant variations across a population, allowing algorithms to classify new samples based on these features. Consequently, transcriptomics can be distinguished from one another with high accuracy.

Prior evidence indicates that peripheral blood transcriptomics change due to predispositions to coronary heart disease. These changes arise primarily from two factors: exposure to known and unknown coronary risk factors, which affect gene expression related to metabolism, immunity, and platelet aggregation, and the inheritance of numerous genome variants that facilitate the development of coronary heart disease and other complex diseases. The trait of a complex disease can be predicted by measuring the polygenic risk score (PRS). Recent studies have confirmed the association of heredity with exposure, integrating both dimensions into a single risk score.

In our preliminary investigations, we focused on refining a peripheral blood sample measurement that incorporates all genetic aspects, including heredity as an expression. We developed a novel approach to analyze a biological sample's transcriptome through convolutional neural networks (deep learning), which demonstrated high precision in classifying samples. In a preliminary study, we identified associations between surrogate markers for cardiovascular disease and gene expression in peripheral blood. For instance, we conducted a clinical study measuring coronary artery calcification (CAC), a reliable surrogate marker of atherosclerotic disease and cardiovascular events, closely related to plaque load, with robust prognostic value and accurately assessed noninvasively through computed tomography (CT). The study revealed clear indications of the clinical utility of our methodology.

The present clinical study aims to identify transcriptomic patterns derived from whole blood samples related to coronary atherotic burden. Additionally, as a secondary analysis, we will explore the algorithm's ability to detect the presence of aortic disease and pro-inflammatory cardiometabolic alterations, such as hepatic steatosis and surrogate markers of coronary inflammation.

4. Background and Rationale

4.1. Cardiovascular disease

Cardiovascular disease (CVD) remains the foremost cause of mortality and morbidity worldwide. In 2019, CVD accounted for approximately 18,600,000 deaths (32.8% of the global total), with 80% of these occurring in low-and-middle-income countries.[1] It is plausible that these figures have risen in the aftermath of the pandemic. Early identification of individuals at an elevated risk of developing asymptomatic

[1] "The Global Burden of Cardiovascular Diseases and Risk Factors." https://pubmed.ncbi.nlm.nih.gov/31727292/. Accessed 10 Apr. 2023.

cardiovascular disease poses a persistent challenge for scientific research and global health authorities. Establishing efficacious and accessible methodologies for CVD detection may yield transformative changes in disease management, substantial reductions in associated healthcare expenditures, and increased life expectancy.

Traditional approaches to identifying individuals with heightened risk of CVD or associated mortality primarily rely on the recognition of singular variables (e.g., diabetes, cholesterol, smoking), a composite of variables (e.g., the Framingham Risk Score), or biomarkers (e.g., Hs-CRP, BNP) linked to an increased likelihood of disease development. Despite these advancements, such methods exhibit limited predictive capacity (Area under the curve 0.60-0.65).[2,3] One potential explanation for this limited efficacy is the numerous processes, both known and unknown, involved in the development of chronic diseases that cannot be wholly encapsulated by analyzing singular or small clusters of predictor variables. Conversely, cardiovascular imaging serves as a valuable tool for early CVD detection (e.g., carotid ultrasound and detection of CAC by CT), but its implementation is hindered by the necessity for costly equipment and specialized personnel. Consequently, the utilization of cardiovascular imaging in clinical practice is limited, and its potential as a population screening instrument remains improbable.

In recent years, substantial progress has been made in identifying genetic variants predisposing individuals to the development of coronary heart disease. Similar to other complex diseases, coronary heart disease exhibits a hereditary component. Traditionally, this hereditary aspect has been discernible in two ways: the manifestation of symptoms at a young age (e.g., myocardial infarction under 45-50 years) and the presence of conspicuous phenotypic traits (e.g., familial hypercholesterolemia) predisposing individuals to cardiovascular events. The latter cases typically result from specific genomic variants (monogenic diseases), representing a minor fraction of the cardiovascular disease patient spectrum. Numerous individuals with a genetic predisposition to CVD inherit a comprehensive set of genomic variants. While this polygenic inheritance is

[2] "Comparison of MESA of and Framingham risk scores in ... - PubMed." https://pubmed.ncbi.nlm.nih.gov/31346632/. Accessed 10 Apr. 2023.

[3] "Multicenter validation of the diagnostic accuracy of a blood-based ...." 5 Oct. 2010, https://pubmed.ncbi.nlm.nih.gov/20921541/. Accessed 10 Apr. 2023.

acknowledged, many genetic variants related to a phenotype lack a definitive causal relationship with the associated pathology. Instead, these variants are only associated with the phenotype as they are part of a haplotype, thereby correlating, either in cis or trans, with other causal variants. The association of coronary heart disease with specific haplotypes is emerging as an effective method for cardiovascular risk prediction.

4.2. Transcriptomics

Cellular DNA safeguards the biological information of a species, transmitting it across generations. However, DNA interacts with the environment in various ways. The most direct and recognized interaction occurs through the conversion of its code via RNA expression, a process known as transcription. The entirety of expressed transcripts at a given moment is termed the transcriptome. Consequently, transcriptomics can be understood as the intersection of DNA's information reserve with the modulation of environmental signals.

Within DNA structure lie units called genes. Until recently, a gene was defined as a DNA sequence transcribed into RNA and subsequently translated into a protein. According to this definition, the human genome comprises approximately 22,000 genes, a number not drastically different from that of a fly (around 14,000), despite the stark phenotypic differences between the two species. However, recent advancements have prompted the redefinition of a gene and its numerical count, which may now be as high as 100,000.[4] Many DNA sequences are transcribed but not translated, collectively referred to as non-coding-RNA. These non-coding genes primarily account for the phenotypic differences between species, individuals, tissues, and cells. Additionally, the genes represented in DNA sequences undergo significant post-transcriptional modifications, such as splicing or circularization, further diversifying potential differences between biological samples. In essence, RNA embodies the dynamic and active component of genetic information, heavily influenced by environmental factors.

[4] "Open questions: How many genes do we have? - BMC Biology." 20 Aug. 2018, https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-018-0564-x. Accessed 10 Apr. 2023.

Technological advancements have facilitated the comprehensive analysis of transcriptomes present in a sample at a specific time. Until a few years ago, RNA sequencing was conducted partially through a technique called microarray; however, it is now predominantly performed using high-throughput RNA sequencing (RNA-Seq). This methodology enables the identification of all RNAs present at a given moment within a cell or tissue. Nevertheless, the vast amount of information generated requires organization and interpretation. Various software tools can align the data with the human genome, identifying individual genes and quantifying their expression levels.

The development of high-throughput sequencing and transcriptomics should be contextualized within the realm of reduced costs. Previously, the cost per sample exceeded one thousand dollars, but the introduction of more advanced equipment has significantly lowered the cost per sample. This decrease in cost is vital for conducting clinical studies involving tens or hundreds of patients. The forthcoming years will witness a continued expansion of RNA-Seq usage in patient studies, with the determination of its clinical utility reliant upon adequate analysis. In this regard, the development of artificial intelligence-based techniques will play a crucial role in facilitating data interpretation.

The integration of transcriptomics with other omics approaches, such as genomics, epigenomics, proteomics, and metabolomics, presents a promising opportunity for a more comprehensive understanding of the complex regulatory networks governing cellular functions and disease processes. By combining data from these diverse sources, researchers can gain deeper insights into the multi-layered interplay between genetic, epigenetic, and environmental factors that contribute to disease development and progression.

One emerging area of interest within the field of transcriptomics is the study of long non-coding RNAs (lncRNAs), which are RNA molecules longer than 200 nucleotides that do not code for proteins. LncRNAs have been implicated in numerous cellular processes, including chromatin remodeling, transcriptional regulation, and post-transcriptional processing. Moreover, lncRNAs have been associated with various human diseases, such as cancer, neurological disorders, and cardiovascular diseases.

Investigating the roles of lncRNAs in these pathologies may yield novel therapeutic targets and diagnostic biomarkers.

Circular RNAs (circRNAs) are a class of non-coding RNA molecules characterized by their unique covalently closed-loop structure, which confers stability and resistance to exonucleases[5]. Although initially regarded as transcriptional byproducts, recent research has demonstrated that circRNAs play crucial roles in various cellular processes, including gene regulation, protein translation, and the modulation of microRNA (miRNA) function through sponge-like interactions. The stability, tissue-specific expression patterns, and conservation across species render circRNAs potential candidates for diagnostic biomarkers and therapeutic targets. Furthermore, circRNAs have been implicated in numerous diseases, including cancer, neurodegenerative disorders, and cardiovascular diseases. A deeper understanding of the functional roles and regulatory mechanisms of circRNAs may lead to the development of novel diagnostic tools and therapeutic strategies for a wide range of human diseases.

The blood microbiome, which refers to the collection of microorganisms present in the circulatory system, has recently emerged as an area of interest for its diagnostic potential. Although the bloodstream has traditionally been considered a sterile environment, advances in next-generation sequencing technologies have revealed the presence of a diverse microbial community within the blood. Alterations in the composition and diversity of the blood microbiome have been associated with various pathological conditions, such as cardiovascular diseases[6], infectious diseases, autoimmune disorders, and cancer. The blood microbiome's accessibility and its potential to reflect systemic changes in response to disease processes make it an attractive candidate for non-invasive diagnostic applications. The development of sensitive, high-throughput technologies to analyze the blood microbiome may facilitate the

[5]"The biogenesis, biology and characterization of circular RNAs - Nature." 8 Aug. 2019, https://www.nature.com/articles/s41576-019-0158-7. Accessed 10 Apr. 2023.

[6] "The blood microbiome and its association to cardiovascular disease ...." 31 Jul. 2022, https://bmccardi ovascdisord.biomedcentral.com/articles/10.1186/s12872-022-02791-7. Accessed 10 Apr. 2023.

identification of novel biomarkers for early disease detection, risk stratification, and monitoring of therapeutic responses.

4.3. Polygenic risk score and its functional derivations

As previously discussed, heredity plays a significant role in some patients with coronary disease. In 2007, Samani et al. reported the association of three haplotypes with early coronary disease.[7] Since then, numerous studies have demonstrated the existence of specific haplotypes for various chronic diseases, including atherosclerosis and coronary heart disease.[8] These haplotypes encompass dozens of genomic variants associated with a phenotype, each having a subtle association with the disease. This minimal association could be a result of the absence of a direct causal relationship, and the existence of the minimal relationship is due to the lack of independent segregation of the alleles (linkage disequilibrium).

Building on the knowledge of the minimal association between variants and the possibility of low-cost, massive DNA sequencing, population studies have facilitated the development of risk scores for the onset of complex diseases based on the summation of minimal inherited risks. Consequently, Polygenic Risk Scores (PRS) have been designed in recent years, which are derived, for example, from the algebraic sum of the odds ratios of each variant. The higher the score, the greater the risk. This approach of aggregating minimal risks has proven effective in discriminating individual risks. As such, a person in the highest-risk decile has the same risk as a patient with familial hypercholesterolemia.[9],[10] The potential use of PRS was described in the context of patients with chest pain. A subanalysis of the PROMISE study found that PRS was independently associated with coronary artery disease in outpatients with stable chest

[7] "Genomewide association analysis of coronary artery disease." 2 Aug. 2007, https://pubmed.ncbi.nlm.nih.gov/17634449/. Accessed 10 Apr. 2023.

[8] "Genetics of Common, Complex Coronary Artery Disease - PubMed." https://pubmed.ncbi.nlm.nih.gov/30901535/. Accessed 10 Apr. 2023.

[9] "Genomic Risk Prediction of Coronary Artery Disease in 480,000 ...." https://discovery.ucl.ac.uk/id/eprint/10057841. Accessed 10 Apr. 2023.

[10] "Clinical utility of polygenic risk scores for coronary artery disease." https://pubmed.ncbi.nlm.nih.gov/34811547/. Accessed 10 Apr. 2023.

pain.[11] Recently, a new study has expanded the knowledge of the association of PRS, identifying 279 loci associated with coronary artery disease.[12] These loci were discovered after analyzing more than 2 million variants sequenced from 9 studies. However, one limitation of these studies is the lack of diversity. When these findings were compared with a Japanese database, a moderate correlation was found with European ancestry. Larger projects from various countries, including Our Future Health in the UK and All of Us in the US, will help develop more robust scores in the near future.

The clinical utility of PRS is currently in the phase of clinical application and may become routine in the coming years. However, PRS only measures the hereditary component of risk. Various researchers have proposed utilizing modified PRS, done by including a component related to gene expression to establish a risk score that also involves the environmental component of risk, which may arise from the association of genetic variants with epigenetic or transcriptome variations.[13] This has led to the concept of Transcriptome-wide Association Studies (TWAS), which measures risk based on the combination of inherited genetic variants and the expression of genes near the same variants.[14,15] Thus, obtaining information from the functional part of the genome has been shown to potentially improve the clinical applications of genomic science.

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

Atherosclerosis is a vascular disease characterized by the accumulation of cholesterol and inflammatory cells in the middle layer of the arteries. The consequences

[11] "Associations of a polygenic risk score with coronary artery disease ...." 21 May. 2022, https://pubmed.ncbi.nlm.nih.gov/35605652/. Accessed 10 Apr. 2023.

[12] "Discovery and systematic characterization of risk variants and genes ...." 6 Dec. 2022, https://www.nature.com/articles/s41588-022-01233-6. Accessed 10 Apr. 2023.

[13] "From GWAS to Function: Using Functional Genomics to Identify the ...." 6 Apr. 2020, https://www.frontiersin.org/articles/10.3389/fgene.2020.00424/full. Accessed 10 Apr. 2023.

[14] "Opportunities and challenges for transcriptome-wide association ...." https://pubmed.ncbi.nlm.nih.gov/30926968/. Accessed 10 Apr. 2023.

[15] "Multi-Omic Strategies for Transcriptome-Wide Association Studies." 8 Mar. 2021, https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009398. Accessed 10 Apr. 2023.

of these deposits are well-established: exposure to lipid-containing inflammatory cells following plaque rupture leads to the majority of acute intravascular thrombotic events, including myocardial infarction, certain types of stroke, and peripheral arterial disease.[16,17] Non-occlusive events can eventually result in low blood flow ischemia.

From the above, it is evident that the complex interaction between inflammatory cells, metabolism, and the vascular wall plays a crucial role in the pathophysiology of atherosclerotic disease. In clinical practice, this interaction is often assessed through surrogate biomarkers of vascular status and the likelihood of adverse outcomes in atherosclerotic disease. These markers can represent metabolic changes (e.g., cholesterol, blood glucose, free fatty acids), inflammatory factors (e.g., C-reactive protein, interleukins, leptin), or coagulation elements (e.g., fibrinogen, PAI-1). Moreover, numerous unidentified markers in peripheral blood may stem from newly described cardiovascular risk factors, such as the intestinal microbiome[18],[19] or air pollution.[20] These markers can be of various molecular types, including proteins, various forms of RNAs, or small molecules.

Among the potential biomarkers mentioned above are RNA molecules. Blood cells are exposed to a wide range of signals from the tissues they perfuse, which can influence circulating cells or even progenitors in the bone marrow, ultimately modifying gene expression. Consequently, different disease states may exhibit distinct sets of differentially expressed genes. Numerous studies have verified the presence of differentially expressed genes in patients with atherosclerotic disease, with alterations in gene expression observed in various stages of atherosclerosis predisposition[21] , prognostic

[16] "Immunity and Inflammation in Atherosclerosis | Circulation Research." 17 Jan. 2019, https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.118.313591. Accessed 10 Apr. 2023.

[17] "Atherosclerosis and inflammation: overview and updates - PubMed." https://pubmed.ncbi.nlm.nih.gov/29930142/. Accessed 10 Apr. 2023.

[18] "Association between the gut microbiome and atherosclerosis - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5815826/. Accessed 10 Apr. 2023.

[19] "The Microbiome and Risk for Atherosclerosis - PubMed." https://pubmed.ncbi.nlm.nih.gov/29800043/. Accessed 10 Apr. 2023.

[20] "Associations between particulate matter air pollution, presence and ...." https://europepmc.org/article/med/32682146. Accessed 10 Apr. 2023.

[21] "Alterations of a Cellular Cholesterol Metabolism Network Are ... - NCBI." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4587646/. Accessed 10 Apr. 2023.

progression[22], or response to statin treatment[23]. Furthermore, the expression of specific gene sets has been associated with clinical variables to predict the severity of atherosclerotic disease[24]. These studies demonstrate the existence of altered gene expression in patients with cardiovascular disease and suggest that these changes may serve as markers for disease and its progression in the future[25,26,27,28].

4.5. Artificial intelligence

Artificial intelligence (AI) has experienced significant growth in recent years, owing to the development of novel algorithms that have substantially enhanced classification capabilities. The potential applications of AI across various fields of human activity have led to its recognition as a pivotal milestone in human evolution. Particularly in medicine, AI has the potential to bring about profound transformations in practice[29].

AI can be characterized as an assembly of algorithms and equations capable of emulating specific cognitive behaviors observed in humans. Although the term implies reasoning capabilities, the current scope of AI is predominantly limited to the human-like (or even super-human) capacity to interpret and classify large volumes of data. This is achieved through the utilization of mathematical algorithms collectively referred to as machine learning. Among these, neural networks have gained prominence in recent years. A neural network consists of numerous parallel and serial linear equations, each

[22] "Low MT-CO1 in Monocytes and Microvesicles Is Associated With ...." https://pubmed.ncbi.nlm.nih.gov/27919931/. Accessed 10 Apr. 2023.

[23] "Intracoronary Imaging, Cholesterol Efflux, and Transcriptomes After ...." 14 Feb. 2017, https://pubmed.ncbi.nlm.nih.gov/27989886/. Accessed 10 Apr. 2023.

[24] "Comparison of MESA of and Framingham risk scores in ... - PubMed." https://pubmed.ncbi.nlm.nih.gov/31346632/. Accessed 10 Apr. 2023.

[25] "The Transcriptomic Toolbox: Resources for Interpreting Large Gene ...." 29 Apr. 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6617151/. Accessed 10 Apr. 2023.

[26] "Transcriptomic Signature of Atherosclerosis in the Peripheral Blood." https://pubmed.ncbi.nlm.nih.gov/27815828/. Accessed 10 Apr. 2023.

[27] "Gene Expression Signatures and the Spectrum of Coronary Artery ...." 19 Jun. 2015, https://link.springer.com/article/10.1007/s12265-015-9640-6. Accessed 10 Apr. 2023.

[28] "Developing Peripheral Blood Gene Expression-Based Diagnostic ...." 25 Jun. 2015, https://link.springer.com/article/10.1007/s12265-015-9641-5. Accessed 10 Apr. 2023.

[29] "High-performance medicine: the convergence of human ... - Nature." 7 Jan. 2019, https://www.nature.com/articles/s41591-018-0300-7. Accessed 10 Apr. 2023.

modified by one or more nonlinear equations, forming an interconnected network of equations that culminate in a nonlinear equation determining probability. This probability is subsequently used for probabilistic classification. While this algorithm has been known for decades, particular advancements in recent years have facilitated its tremendous expansion, leading to the emergence of deep learning[30]. These advancements are primarily twofold: firstly, the development of a self-correcting capability for the network in the 1980s through the implementation of an algorithm called backpropagation[31],[32]. This algorithm enables the neural network to adjust in successive cycles, progressively enhancing its predictive capacity through continuous training. The second contributing factor to the proliferation of deep learning is the advancement of hardware capable of performing a vast number of simple calculations (linear regressions) in a short time, primarily through the adaptation of Graphical Processing Units (GPUs) for AI applications. These two developments have led to the remarkable expansion of deep learning, with an increasing presence in everyday processes.

4.6. Use of artificial intelligence in medicine

In recent years, numerous examples have emerged illustrating the potential of AI in medicine. A prominent study published in 2017 demonstrated an algorithm's ability to differentiate malignant skin lesions from photographs, including various forms of skin cancer, with higher accuracy than a group of dermatologists[33]. Another equally significant study showcased the high capacity of deep learning (DL) to identify the presence of tumor cells in histopathological images of axillary lymph nodes in breast cancer patients[34]. These instances exemplify the impact of advancements in AI: a trained algorithm can autonomously detect abnormalities in images, irrespective of a physician's involvement.

[30] "Deep learning | Nature." 27 May. 2015, https://www.nature.com/articles/nature14539. Accessed 10 Apr. 2023.

[31] "Learning representations by back-propagating errors - Nature." 9 Oct. 1986, https://www.nature.com/articles/323533a0. Accessed 10 Apr. 2023.

[32] "Improving generalization performance using double backpropagation." https://ieeexplore.ieee.org/document/165600. Accessed 10 Apr. 2023.

[33] "Dermatologist-level classification of skin cancer with deep neural ...." https://pubmed.ncbi.nlm.nih.gov/28117445/. Accessed 10 Apr. 2023.

[34] "Diagnostic Assessment of Deep Learning Algorithms for Detection of ...." 12 Dec. 2017, https://pubmed.ncbi.nlm.nih.gov/29234806/. Accessed 10 Apr. 2023.

The vast diagnostic potential of these algorithms has provided a strong impetus for further research in this domain.

While the application of AI in medicine remains in its early stages and primarily in the research phase, potential uses can be categorized into several areas:

- Radiology: AI algorithms can process imaging studies, learning characteristics associated with abnormal defects. With an increasing number of potential AI applications in this field, some have already been approved for clinical use. In the near future, many common radiological images will likely be pre-scanned by algorithms to identify abnormalities, streamlining physicians' workload by allowing them to focus only on images containing pathological alterations. Various studies have demonstrated the utility of DL in multiple areas of cardiovascular medicine, from risk stratification and triage tools to the automatic evaluation of ventricular volumes in magnetic resonance imaging and coronary calcium scores by CT.
- Body images: Due to DL's success in image processing, promising applications are also anticipated in this area. Examples include the interpretation of skin lesions and the analysis of fundus images to potentially predict cardiovascular risk. Moreover, neural networks can determine whether an eye fundus image belongs to a male or female with approximately 98% accuracy, indicating their capacity to extract imperceptible features from an image. The precise characteristics defining these differences remain unclear but may be related to arterial thickness or other subtle variations[35].
- Continuous signals: AI can be utilized to interpret data from continuous acquisition signals, such as electrocardiograms or electroencephalograms, using DL algorithms[36,37].

[35] "Prediction of cardiovascular risk factors from retinal fundus ...." https://pubmed.ncbi.nlm.nih.gov/31015713/. Accessed 10 Apr. 2023.

[36] "Cardiologist-level arrhythmia detection and classification in ... - Nature." 7 Jan. 2019, https://www.nature.com/articles/s41591-018-0268-3. Accessed 10 Apr. 2023.

[37] "Deep learning for electroencephalogram (EEG) classification tasks." https://iopscience.iop.org/article/10.1088/1741-2552/ab0ab5. Accessed 10 Apr. 2023.

- Language processing: DL is revolutionizing language interpretation, with applications including the automated analysis of medical records and patient vocabulary. Numerous reports have identified emotions, including depressive states, based on a subject's language use[38].

Additionally, Generative Pre-trained Transformers (GPT) and other large language models (LLMs) hold considerable potential in medicine[39]. These models can be employed for tasks such as:

- Summarizing and extracting relevant information from lengthy medical documents, aiding physicians in quickly understanding the most critical aspects.
- Assisting in the creation of accurate and coherent patient education materials, thereby improving patients' comprehension of their diagnoses, treatments, and overall health management.
- Facilitating the development of personalized treatment plans by analyzing patient medical histories and suggesting suitable interventions based on relevant research findings and best practices.
- Enhancing the efficiency of electronic health record systems through natural language processing, enabling healthcare professionals to retrieve relevant information more quickly and accurately.
- Supporting medical research by reviewing existing literature, identifying gaps in knowledge, and generating new hypotheses for further investigation.

The application of GPT and LLMs in medicine promises to enhance diagnostic accuracy, streamline workflows, and ultimately contribute to improved patient care and outcomes. However, it is crucial to ensure data security and privacy and address ethical concerns surrounding the use of AI in healthcare.

[38] "A Novel Method to Detect Functional microRNA Regulatory Modules ...." https://pubmed.ncbi.nlm.nih.gov/27295638/. Accessed 10 Apr. 2023.

[39] "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine." https://www.nejm.org/doi/10.1056/NEJMsr2214184. Accessed 10 Apr. 2023.

5. Preliminary observations

5.1. In silico developments

In our laboratory, we have developed a comprehensive methodology for transcriptome analysis that combines extensive bioinformatic processing of biological samples with deep learning algorithms for interpretation. This approach enables the incorporation of all genes read in a sequence, performing classifications using a supervised method. Our algorithm has demonstrated robustness and high accuracy in preliminary tests.

The development consists of a pipeline that analyzes the RNA-seq sample in its crudest state, including coding and non-coding RNAs, circular RNAs, and the blood microbiome. The algorithm is trained on a cohort of previously classified samples and later validated using 10% to 20% of the samples.

To assess the algorithm's efficiency, we analyzed different RNA-seq databases available in digital repositories. Initially, we used the GTEx database, containing 11,866 sequences from 30 different tissues obtained postmortem from dozens of individuals. Our algorithm achieved 99.1% accuracy in correct prediction (Figure 1), with only a few errors corresponding to potentially similar tissues (e.g., breast and fatty tissue).

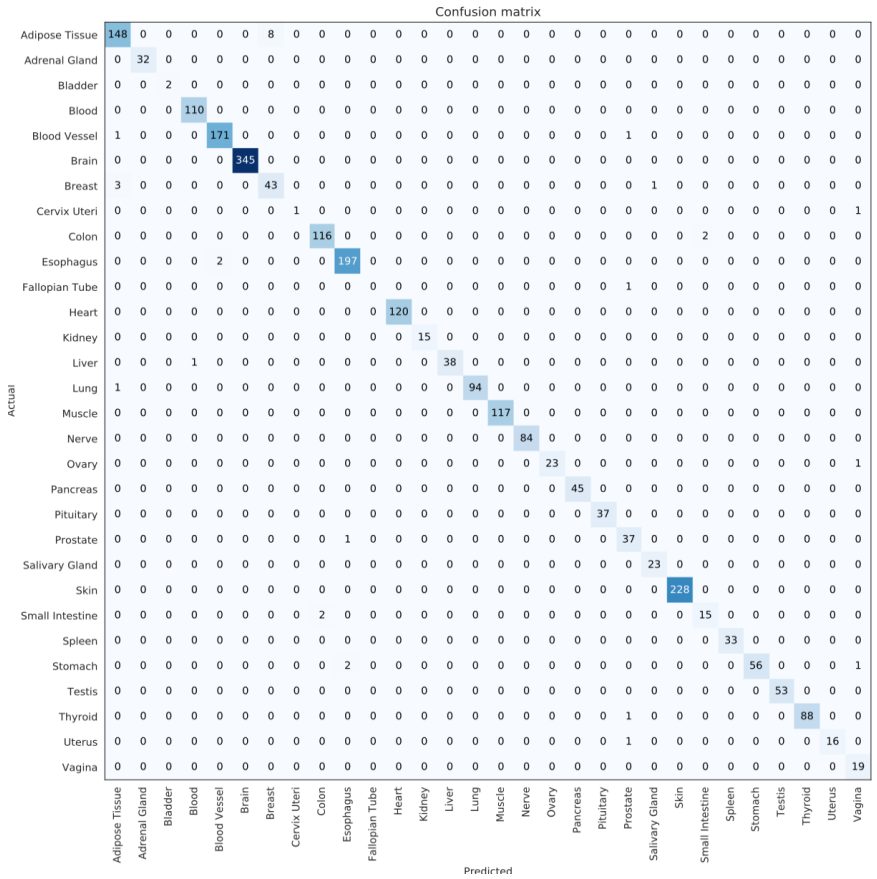

Figure 1: Confusion matrix for normal tissue samples. The graph shows the intersection of the algorithm's prediction and the actual sample. Thus, the descending central diagonal from left to right shows the coincidence between the real and the predicted. Outside this diagonal, the numbers represent the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

We also analyzed the Genomic Data Commons (GDC) database, containing RNA sequencing of biopsies from 34 different tumors with 10,000 samples. The algorithm's predictive capacity was again high, reaching 97.5% accuracy (Figure 2). These results demonstrate the algorithm's excellent classification capability for multiple samples.

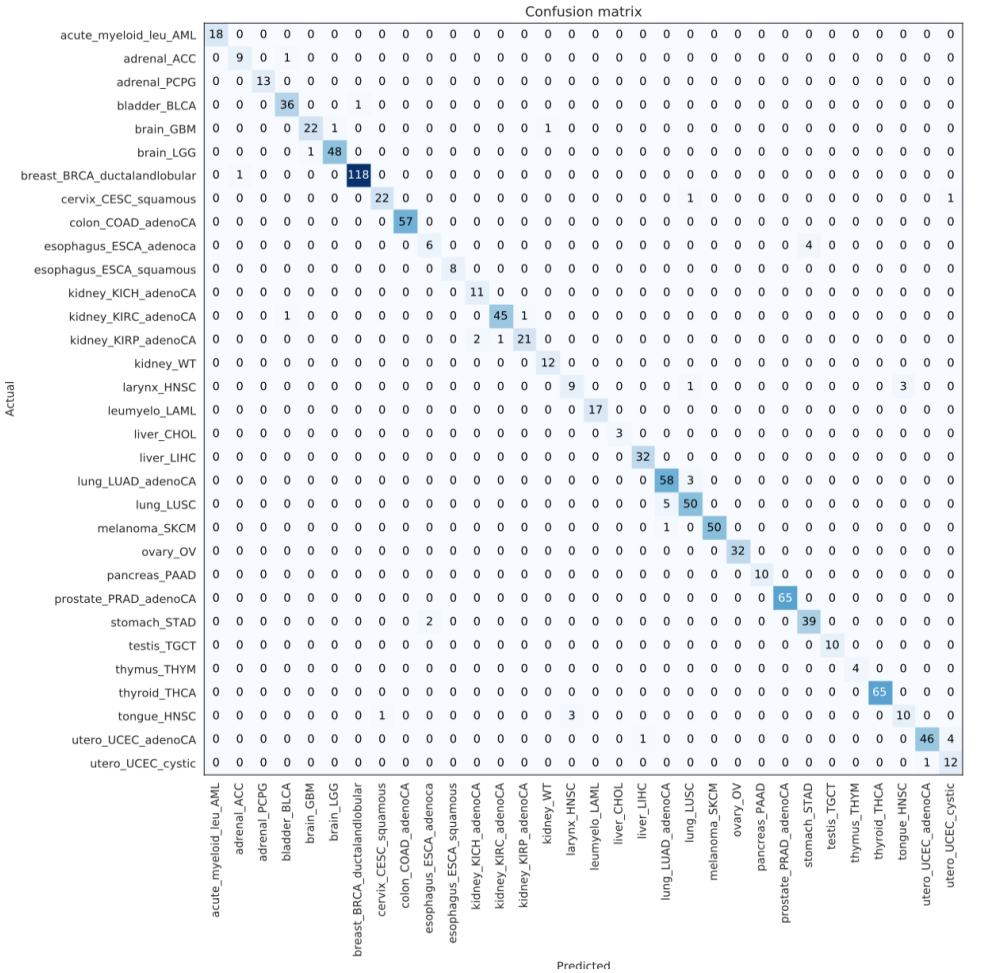

Figure 2: Confusion matrix for tumor tissue samples. The graph has the same meaning as in the previous figure. Again, the number of mistakes outside the diagonal line is very low.

A sensitivity analysis of the algorithm was performed using 100 random samples from brain tumors in the GDC database. The algorithm achieved 100% precision (ROC of 1) and exhibited efficient behavior with a low number of clinical cases, reflecting the fundamental differences in gene expression between the two groups of tumors.

The aforementioned results were obtained from solid tissue RNA-Seq samples (biopsies). It is worth noting that blood cells may exhibit transcriptomic alterations in the presence of solid tumors or systemic diseases. Therefore, we analyzed different datasets

in various contexts, including samples obtained by Best et al[40] from patients with different solid tumors. The algorithm showed an AUC of 1 for detecting the presence of tumors in all cases.

We also analyzed samples from Cassetta et al on monocytes from patients with breast or endometrial cancer compared to healthy subjects[41], and the algorithm achieved 100% prediction capacity. Moreover, we analyzed the dataset generated by Scharer et al on patients with lupus compared to healthy subjects, and the classification was excellent (AUC of 1)[42].

In conclusion, our pipeline and the analyses performed demonstrate the ability to classify transcriptomics samples accurately. The algorithm is robust and extremely sensitive, primarily due to the large number of gene variants considered (approximately 200,000) and the unprecedented capacity of neural networks to interpret this vast amount of information. These findings lead us to propose that our pipeline may be useful in detecting tumor pathologies in peripheral blood transcriptomics.

5.2. Preliminary studies

We have previously developed a clinical protocol aimed at refining the methodology that will be employed in this study. The initial objective was to establish a reliable process for blood collection, RNA isolation, library generation for sequencing, and RNA sequencing. The preliminary results of these steps, conducted on 540 samples from three different studies, are described below.

First, we established a robust methodology for obtaining RNA samples from peripheral whole blood without cell separation. Samples were collected in PaxGene RNA tubes through antecubital vein puncture, and no adverse effects were observed in this procedure. The samples were stored at -20°C until RNA extraction, which was carried out

[40] "RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan ...." https://pubmed.ncbi.nlm.nih.gov/26525104/. Accessed 10 Apr. 2023.

[41] "Human Tumor-Associated Macrophage and Monocyte ...." 15 Apr. 2019, https://scholars.duke.edu/individual/pub1379736. Accessed 10 Apr. 2023.

[42] "Epigenetic programming underpins B cell dysfunction in human SLE." https://pubmed.ncbi.nlm.nih.gov/31263277/. Accessed 10 Apr. 2023.

using a specific kit from Thermo-Fischer. This approach allowed us to consistently obtain RNA samples in sufficient quantities for sequencing in all cases.

Next, we constructed the library, a process in which the RNA is converted into cDNA (complementary DNA). The RNA corresponding to ribosomes and globin is then extracted, fragmented, and adapters are added to the ends. The library construction proved robust for all analyzed samples.

Subsequently, we performed RNA sequencing at a depth of 100 million reads. Between 20,000 and 25,000 genes expressed in peripheral blood were detected across all samples (n=540). In a subset of patients, computed tomography was carried out to determine the degree of calcification in the coronary arteries (Coronary Calcium Score, CCS). We analyzed gene expression in peripheral blood and identified genes related to the degree of coronary calcification (Figure 3).

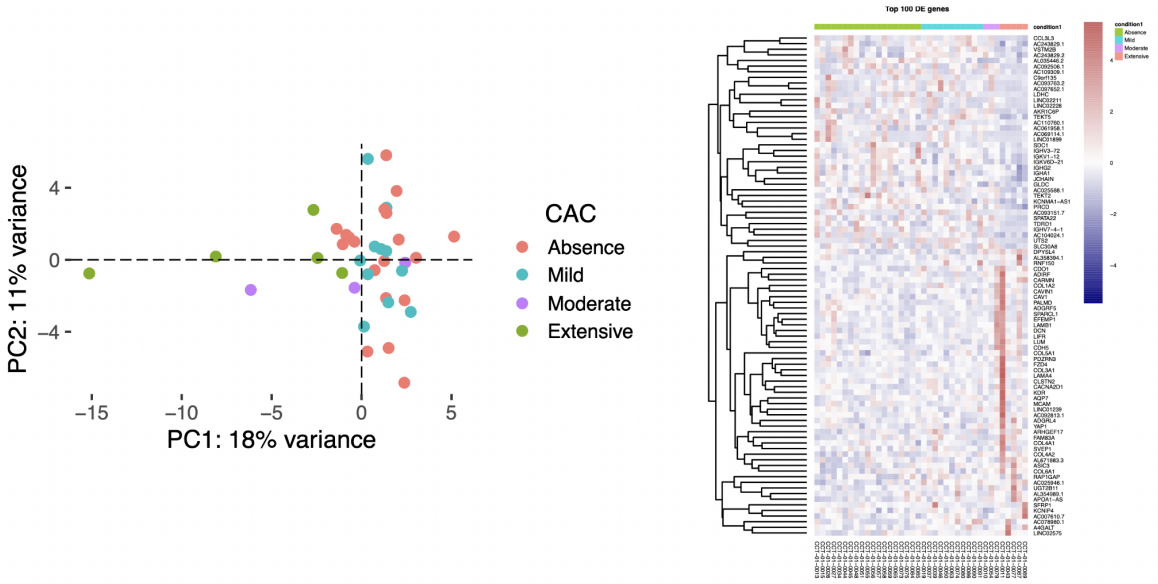

Figure 3: Preliminary results. We analyzed 38 samples from patients who had coronary calcium measured. The left panel shows the result of a PCA analysis where a moderate but significant reduction in dimensionality is found based on the extent of coronary calcium. The heatmap shows the differentially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

6. Study Objectives

6.1 Primary Objective

The main objective of this study is to investigate the diagnostic accuracy of whole-blood transcriptome analyzed by artificial intelligence in detecting the presence and extent of coronary atherosclerosis in stable patients with suspected or known coronary artery disease (CAD), using coronary computed tomography angiography (CCTA) images.

Specific aims:

I) To classify patient transcriptomes based on the presence and extent of the following CCTA outcomes:

- Coronary calcium score
- Total coronary plaque burden
- Degree of coronary stenosis
- Coronary high-risk plaques (low-attenuation, positive remodeling; spotty calcification, or napkin-ring sign)

II) To train artificial intelligence algorithms using information from the transcriptome to recognize the outcomes of interest in the training set.

III) To explore the diagnostic accuracy (sensitivity and specificity) of the method in detecting the presence, absence, or extent of the outcomes of interest in the validation/test set.

6.2 Secondary Objectives:

The secondary objectives of this study are as follows:

6.2.1 To explore the diagnostic accuracy of whole-blood transcriptome in relation to other outcomes of interest provided by CCTA images, including:

- Presence and extent of aortic calcification
- Hepatic steatosis
- Flow limiting artery obstruction (coronary fractional flow reserve)
- Coronary artery inflammation (perivascular fat attenuation index)
- Total epicardial adipose tissue volume

6.2.2 To calculate the polygenic risk score (PRS) from the transcriptome and explore its association with the previously described outcomes of interest.

6.2.3 To explore the differences in transcriptomics of peripheral blood between patients who did and did not develop a cardiovascular event during follow-up.

7. Methods

7.1. Study Design

This will be a prospective observational study. A convenience sample will be carried out to include 200 patients who attend the ENERI Medical Institute, La Sagrada Familia Clinic and Sanatorio Mendez with a clinical indication to be evaluated by a CCTA due to suspected CAD or known CAD. 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 mucose will be taken. Subsequently, the images obtained from the clinically indicated CCTA will be assessed to explore the outcomes of interest. At the end of patient enrollment, biological samples will be sequenced for in silico evaluation of the results. Finally, a 5-year follow-up will be carried out via telephone or email contact to collect data on the incidence of fatal and non-fatal cardiovascular events.

8.Study Population

8.1. Inclusion Criteria

- Men and women between 18 and 75 years, with a clinical indication to be evaluated by a CCTA due to suspected or known CAD.
- Signature of informed consent.

8.2. Exclusion criteria.

- Previously known chronic renal or hepatic insufficiency.
- Active chronic lung disease, defined as: exacerbated asthma, exacerbated COPD, or pulmonary fibrosis.
- Myocardial infarction, unstable angina, cerebrovascular accident, or vascular interventions (any territory) in the past 6 months.
- Implantation of drug-eluting stents in the last 12 months.
- Revascularization surgery (By-pass)
- Indication of Angio-CT for congenital heart disease or Transcatheter Aortic Valve Replacement (TAVI)
- Presence/ diagnosis of heart failure symptoms or signs (ie. dyspnea, asthenia, edema) with objective evidence of pulmonary or systemic congestion at rest or with exercise in the last 6 months[43].
- Left ventricular ejection fraction <50% confirmed by objective diagnostic methods (eg doppler echocardiogram).
- Severe valvulopathies, confirmed by objective diagnostic methods (eg doppler echocardiogram).
- Uncontrolled hyper or hypothyroidism.

[43] "2013 ACCF/AHA guideline for the management of heart failure." https://pubmed.ncbi.nlm.nih.gov/23741058/. Accessed 10 Apr. 2023.

- Suprarenal 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.
- Diagnosis of active autoimmune disease or any pathology under immunosuppressive treatment.
- Ongoing pregnancy, postpartum period of less than 12 months or breastfeeding.
- Other serious diseases with an estimated life expectancy of less than 12 months (according to the investigator's opinion).
- 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.
- Pacemaker/Cardio-Defibrillator Implantation.
- Prior to the study: heart rate > 70 l/m or atrial fibrillation or frequent extrasystoles that in the opinion of the specialist physician will affect the quality of the cardiovascular imaging.

9. Procedures

9.1. Eligibility and informed consent

Patients referred to CCTA scans with a medical indication due to suspected CAD or known CAD, will be recruited at Clínica Sagrada Familia and Sanatorio Mendez in Buenos Aires. 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 patients' questions, and once they have expressed that they agree to donate blood samples, mucose DNA cells via oral swab, associated data and the CCTA resulting image to the study, they will be formally invited to sign the informed consent.

9.2. 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 designed for this study:

9.2.1 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.

9.2.2 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 fast for 8 hours prior to the study. 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, C-reactive protein and leukocyte formula. The primary tube for clinical laboratory analysis will be sent for analysis to local laboratories which are: Stamboulián and Sanatorio Méndez.

9.2.3 Oral mucose swab (mouth swab)

To obtain the sample, the professional will rub a buccal swab against the inside of the patient's cheeks and gums, and place it in a dry sterile 1.5ml eppendorf-type tube with lysis buffer; once labeled with the participant's ID, it will be placed 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.

9.2.4 Coronary Computed Tomography Angiography

Participants' donated CCTA images and associated data will be collected and anonymised at the site for further analysis.

CCTA acquisition protocols will follow each site's standard procedures. The use of any medications involved in routine CCTA examinations such as beta blockers (in the case of a heart rate higher than 65 beats per minute) or sublingual nitroglycerin will not be encouraged, the decision being left at the discretion of the physician involved in the acquisition. CTCA images will be transferred to an AWS server for independent data analysis.

9.3. 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 kit TruSeq-RiboZero-GlobinFree (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.

9.4. 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 a *pipeline* analysis which 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:

- Linear RNA identification, including mRNA and non-coding RNAs.
- circRNA (circular RNA)
- Alternative splicing variant identification
- Blood microbiome
- Variant calling for Polygenic risk score calculation (from RNA).

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 accordingly.

Once the bioinformatic analyses are done, we will use that information to train machine and deep learning models to predict those patients with coronary artery disease. The following ML/DL approaches will be used:

- Classical machine learning algorithms, including xgboost.
- Convoluted Neural Networks (CNN) through our patented technology of gene expression conversion.
- Graph neural networks.
- Natural processing language by using reads.

All these bioinformatic and artificial intelligence procedures will be 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 the training of the neural network.

Data will be stored and analyzed at the cloud provider at the moment of analysis (currently AWS or Google Cloud) under strict cybersecurity measures.

9.5. Analysis of Coronary Computed Tomography Angiography

Coronary Calcium scoring will be analyzed using dedicated software (HeartBeat-CS, Philips Healthcare, Best, The Netherlands) by an experienced observer blinded to the clinical characteristics, laboratory, and transcriptome results. 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 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 the area ranging from the bifurcation of the pulmonary artery as the upper limit to the diaphragm as the lower limit.

CCTA images will be anonymised locally and transferred to a core laboratory located in Oxford, UK, for analysis of secondary outcomes by investigators who will be unaware of population demographics and outcomes. The presence of obstructive CAD will be defined as the presence of at least one lesion causing a luminal stenosis of 50% or larger. The extent of coronary artery disease using the modified Duke coronary artery disease index, as follows (Group 1: <50% stenosis; Group 2: ≥2 mild stenosis with proximal CAD in 1 artery or 1 moderate stenosis; Group 3: 2 moderate stenosis or 1 severe stenosis; Group 4: 3 moderate stenosis, 2 severe stenosis, or severe stenosis in the proximal left anterior descending (LAD) artery.; Group 5: 3 severe stenosis or 2 severe stenosis in the proximal LAD; Group 6: ≥50% stenosis in left main coronary artery). The presence of high-risk plaque features (spotty calcification, low-attenuation plaque, positive remodeling, and napkin-ring sign) will be explored. The total epicardial adipose tissue volume will be calculated in a semi-automated manner by tracking the contour of the pericardium from

the level of the pulmonary artery bifurcation to the apex of the heart at the most caudal end. To measure the perivascular FAI, we will trace the proximal 40-mm segments of all three major epicardial coronary vessels (right coronary artery, left anterior descending artery, and left circumflex artery) and defined respective perivascular fat as the adipose tissue within a radial distance from the outer vessel wall equal to the diameter of the vessel, as described and validated previously. To avoid the effects of the aortic wall, we excluded the most proximal 10 mm of the right coronary artery and analyzed the proximal 10–50 mm of the vessel, as described previously. In the left anterior descending artery and left circumflex artery, we will analyze the proximal 40 mm of each vessel.

As follows, 3-dimensional plaque quantification of the whole vessel will be performed (10,13) using semi-automated plaque analysis software. All IDs will undergo 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.

9.6. Storage of biological samples

The remnant 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 the laboratory. 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.

9.7. Patient follow-up

Patients included in the study will be contacted annually by phone call or email 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.

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

The study will assess the following incident events:

- Death from any cause.
- Fatal and non-fatal atherosclerotic related events: AMI, angina, cerebrovascular accident (“stroke”) or transient ischemic attack.
- Hospitalized (emergency room or admitted) Heart failure event with preserved or impaired ventricular function.
- Vascular revascularization: performance of myocardial revascularization surgery, coronary angioplasty or peripheral vascular angioplasty.
- Cancer: new diagnosis of any type of cancer.

9.7.2. Verification and adjudication of events during follow-up

When an event of interest is reported during study follow-up, the "phase of verification” will be triggered 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, GRG). RP and GRG will analyze the documentation sent for the case separately and will adjudicate the event. If there is no agreement, SGM will make the final adjudication.

10. Guarantee and quality control of the data

10.1. Confidentiality of the data

The data of the patients will be collected anonymously from the moment of signing the informed consent. To do this, an ID code representative 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 AWS which is in compliance with the Health Insurance Portability and Accountability Act (HIPAA).

10.2. Data quality assurance and control

a. Data quality assurance and storage

This study will implement a standardized data collection process, since trained interviewers will ensure data collection uniformly across all respondents, properly fill out the forms and handle sensitive information.

Baseline data entry and storage will be performed using the 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 users 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 personnel in charge of entering the data will be trained in the use of RedCap for this purpose.

The study data corresponding to the Follow-up stage will be collected on paper-based forms. The forms will contain validated questions usually used in previous clinical studies.

At the sites, the anonymised paper-based forms will be scanned by trained personnel and stored in a secure location to protect sensitive information from unauthorized access. All anonymized digital forms, CCTA images, genetic data and laboratory reports will be encrypted and stored on a central outsourced AWS server. The access 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.

A random quality check will be implemented on a sample of completed forms for accuracy and completeness to ensure that the data collected is reliable.

The study data will then be stored on a central outsourced server (AWS).

b. Data Entry

Trained and certified personnel will enter the digital data in an Excel format document located in the AWS server.

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 Excel.
- Query generation 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.

Once data has been entered into a digital system, the physical forms will be properly filed to protect sensitive information.

c. Data security control: auditing and monitoring

Monitor and log activity within the AWS environment will be conducted regularly reviewing logs to detect unauthorized access or suspicious activity.

Regular data backup and disaster recovery will be automatically implemented using AWS Backup, in order to ensure the availability and integrity of study data. All the security processes will be performed by personnel certified in data security best practices, capable of handling sensitive information, how to identify phishing attacks, and how to report security incidents. Regular risk assessments and vulnerability scans will be conducted to identify potential threats and vulnerabilities in the AWS environment in order to implement the necessary controls to mitigate the risk.

d. Elaboration of the manual of procedures

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

e. 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, blood samples management, data management, management of an adverse event, follow-up, and measurement procedures, data entry and data security. Those who have acquired the expected skills will be certified to carry out the study procedures. Staff will be retrained on a regular basis.

f. Site Monitoring Visits and Audits

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

10.3. Adverse event reporting

Serious and non-serious events (expected or unexpected) will be reported to the IRB within specific timelines depending on the severity of the event. Serious adverse events will be reported within 15 calendar days of receipt, whereas non-serious adverse events will be reported on a periodic basis.

All the adverse event forms will be kept in the corresponding study records of adverse events, including any follow-up actions taken, for a period of time as specified by the regulatory agency.

10.4. 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. *In silico* studies carried out by this group of researchers and the preliminary clinical studies carried out previously showed that at least a sample of 150 patients is necessary for the initial training of the artificial intelligence algorithm that detects transcriptional differences between patients with or without a solid tumor. (for example, breast tumor). Therefore, 200 patients will be included based on the minimum number necessary to perform the initial training of an artificial intelligence algorithm.

10.5. 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 accuracy 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.

10.6. Control of good clinical practices

The proposed study 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))[44].

[44] "Efficacy Guidelines - ICH." https://www.ich.org/page/efficacy-guidelines. Accessed 10 Apr. 2023.

All study investigators have been certified in GCP guidelines. New study personnel will also be required to obtain this certification.

10.7. Protocol Deviations

Investigators in this study will strictly monitor compliance with good clinical practices. 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.

10.8. 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.

10.9. 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 LTD.

10.10. Study Organization

10.10.1. Direction of the Study

The principal investigator of the project will be Rosana Poggio, while Gastón Rodríguez Granillo and Santiago Miriuka will be the co-principal investigators. Dr. Gastón Rodríguez Granillo will also be the principal investigator of La Sagrada Familia, will provide logistical support and will be responsible for all on-site study execution. Dr. Guillermo Naum will be the principal investigator of Diagnóstico MEDITER / Sanatorio Méndez, and will provide logistical support and be responsible for all on-site study implementation. A

CONICET fellow will be the coordinator of the logistics and processing of the RNA and DNA biological samples. Postdoctoral fellow Alejandro La Greca will be in charge of the bioinformatics analysis. María Olivera will be the general coordinator and study monitor.

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.

10.11. Financing

The funding to cover the processes and materials necessary to carry out this study will be provided entirely by MultiplAI Health LTD (sponsor) and does not imply any expense for the OSBA or for Sanatorio Méndez or Clínica La Sagrada Familia; therefore:

Compensation to the participating professionals: $ 0.00

Laboratory determinations: $0.00

Genetic analysis (sequencing): $0.00

Data upload and analysis: $0.00

10.13 Study Timeline

	Baseline (Months)												Follow-up (Years)				
	1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	II°	12°	1°	2°	3°	4°	5°
Enrollment																	

Baseline phase queries resolution																	
Follow-up																	
Follow-up phase queries resolution																	

The schedule of the present study will begin to be implemented once the study has been approved by the corresponding ethics committee.