



**Università
di Brescia**

DIPARTIMENTO DI SCIENZE CLINICHE E SPERIMENTALI

DOTTORATO DI RICERCA IN INTELLIGENZA ARTIFICIALE IN MEDICINA E INNOVAZIONE
NELLA RICERCA CLINICA E METODOLOGICA

settore scientifico disciplinare: MED/11

CICLO XXXVIII

**Applications of Artificial Intelligence in Cardiology: Methodological Advances
and Three Use Cases**

DOTTORANDO: Mauro Massussi

Primo SUPERVISORE: Prof. Marco Metra

Contents

Abstract	pag. 3
1. General Introduction.....	pag. 6
1.1. Key AI application in Healthcare and thesis structure	
2. Use Case 1: Deep Learning for Predicting Cardiovascular Outcomes from Chest X-Ray Images.....	pag. 8
2.1. Background and Rationale	
2.2. Deep Learning for Diagnostic Triage of Significant Coronary Artery Disease	
2.3. Long-Term Mortality Prediction using Deep Learning on Chest X-Rays	
3. Use Case 2: Self-Supervised Learning for ECG Interpretation (HUBERT-ECG Foundation Model)	pag. 41
3.1. Rationale and Innovation: The Foundation Model Paradigm for ECG Interpretation	
3.2. HuBERT-ECG: Self-Supervised Learning for Scalable, Multitask ECG Diagnosis and Prediction	
4. Use Case 3: The Dorian Gray Project – Linking Cardiovascular Disease and Cognitive Decline through AI	pag. 71
4.1. The DORIAN GRAY Project: Architectural Design for a Cardiovascular Digital Twin	
4.2. Scientific Rationale: The Heart–Brain Axis and Disease Phenomenology	
4.3. Data Specification and AI Grounding: Methodological framework	
4.4. Clinical Trial Design for AI System Validation: the COGIT-HF protocol	
5. Conclusions.....	pag. 100
5.1. Main Contributions of This PhD	
5.2. Broader Implications and Recommendations	

Abstract

Introduction:

Artificial intelligence (AI) is transforming cardiology, enabling a shift from reactive to predictive and personalized medicine. This thesis demonstrates the methodological potential and clinical relevance of AI across three distinct use cases in cardiovascular health, focusing on: opportunistic screening, foundation models for generalizable diagnosis, and digital twin technology for personalized intervention.

Methods and Results:

Opportunistic Screening (Use Case 1): Deep Learning models were applied to routine Chest X-Ray (CXR) images of patients referred for suspected Coronary Artery Disease (CAD). The model accurately predicted the presence of significant CAD for diagnostic triage (AUC 0.77 with angina) and effectively stratified long-term mortality risk (AUC 0.81 when combined with age and CAD status). This work validates CXR as a low-cost, scalable source of prognostic biomarkers.

Foundation Models (Use Case 2): We developed HuBERT-ECG, a self-supervised foundation model, pre-trained on an unprecedented 9.1 million 12-lead ECGs. HuBERT-ECG demonstrated robust generalizability across 17 downstream tasks and achieved AUROCs >0.90 for 140 out of 164 cardiovascular conditions. Crucially, the model leveraged inter-relationship learning on the Cardio-Learning dataset, achieving a dramatic improvement in prediction for complex structural heart disease, such as Heart Failure (HF), where ECG signals are conventionally subtle (AUROC on HF 0.9674 ± 0.0566).

Digital Twin Architecture (Use Case 3): We conceptualized and architecturally designed the DORIAN GRAY Project (funded by Horizon Europe). This project employs a Digital Twin framework and a cascading AI model suite to link Cardiovascular Disease (CVD) and Mild Cognitive Impairment (MCI) through causal inference and personalized risk stratification. The solution utilizes a highly adaptive Avatar-Based Coaching Exergame (ABCE), validated through the dedicated COGIT-HF randomized trial protocol, representing a novel design for continuous, real-time digital intervention.

Conclusions: This thesis successfully demonstrated three scalable and high-impact applications of AI in cardiology. The work advances AI methodology across image analysis, time-series foundation modeling, and complex digital health architectural design. These contributions offer tangible pathways for developing robust, personalized, and efficient tools that enhance clinical decision-making and promote equitable cardiovascular and cognitive health globally

Sinossi in lingua Italiana

Introduzione

L'intelligenza artificiale (IA) sta trasformando la cardiologia, favorendo il passaggio da una medicina reattiva a un approccio predittivo e personalizzato. Questa tesi illustra il potenziale metodologico e l'impatto clinico dell'IA attraverso tre distinti casi d'uso nell'ambito della salute cardiovascolare, focalizzati su: screening opportunistico, modelli fondativi per una diagnosi generalizzabile e gemelli digitali per interventi personalizzati.

Metodi e Risultati

Screening Opportunistico: sono stati applicati modelli di deep learning alle radiografie del torace (CXR) eseguite routinariamente nei pazienti valutati per sospetta coronaropatia (CAD). Il modello è risultato in grado di predire con buona accuratezza la presenza di CAD significativa ai fini del triage diagnostico (AUC 0.77 nei pazienti con angina) e di stratificare efficacemente il rischio di mortalità a lungo termine (AUC 0.81 combinando età e stato CAD). Lo studio dimostra che la CXR può rappresentare una fonte scalabile e a basso costo di biomarcatori prognostici.

Modelli Fondativi (Foundational Models): è stato sviluppato HuBERT-ECG, un modello fondativo auto-supervisionato pre-addestrato su un dataset senza precedenti composto da 9,1 milioni di ECG a 12 derivazioni. HuBERT-ECG ha mostrato un'elevata capacità di generalizzazione in 17 compiti successivi e ha ottenuto un AUROC >0.90 per 140 su 164 condizioni cardiovascolari. In particolare, sfruttando le relazioni interne del dataset Cardio-Learning, ha migliorato significativamente la predizione di patologie cardiache strutturali complesse, come lo scompenso cardiaco (HF), per le quali l'ECG fornisce segnali tradizionalmente poco specifici (AUROC per HF 0.9674 ± 0.0566).

Architettura del Gemello Digitale: è stata definita l'architettura del progetto DORIAN GRAY, finanziato da Horizon Europe, basata su un framework di Gemello Digitale e su una serie di modelli di IA integrati. L'obiettivo è collegare malattia cardiovascolare e lieve decadimento cognitivo tramite inferenza causale e stratificazione personalizzata del rischio. La piattaforma incorpora un Avatar-Based Coaching Exergame (ABCE) ad elevato grado di adattività, validato tramite il trial randomizzato dedicato COGIT-HF, proponendo un approccio innovativo all'intervento digitale continuo e in tempo reale.

Conclusioni

La tesi dimostra tre applicazioni scalabili e ad alto impatto dell'IA in cardiologia. I risultati avanzano lo stato dell'arte nell'analisi di immagini, nella costruzione di modelli fondativi per serie temporali e

nella progettazione di architetture complesse per la sanità digitale. Nel loro insieme, questi contributi delineano percorsi concreti verso strumenti clinici robusti, personalizzati ed efficienti, finalizzati a migliorare il processo decisionale e a promuovere equità nella salute cardiovascolare e cognitiva su scala globale

1. General Introduction

1.1. Key Artificial Intelligence Applications in Healthcare and Thesis Structure

Artificial intelligence (AI) is reshaping modern medicine in ways that extend far beyond the adoption of new computational tools. Its integration into healthcare represents a profound transformation in how we diagnose disease, make clinical decisions, and conceptualize human health. Rather than replacing clinicians, AI is emerging as a powerful ally—one that enhances precision, anticipates risk, and supports more efficient and equitable care delivery¹.

At its core, AI enables a shift from reactive medicine to predictive and personalized healthcare. By uncovering subtle patterns in biomedical data, AI systems can forecast disease trajectories, optimize treatment strategies, and support clinical workflows with an accuracy that continues to approach, and sometimes surpass, traditional methods. This paradigm shift, however, comes with important challenges. Ensuring the reliability, transparency, and safety of AI models remains essential for clinical adoption, particularly when algorithmic recommendations may influence high-stakes decisions. Ethical questions regarding autonomy, accountability, and explainability cannot be treated as secondary considerations: they are foundational to building trust in AI-assisted medicine².

The rapid evolution of AI over the past decade highlights not only unprecedented opportunities, but also important responsibilities. Issues of privacy, security, fairness, and equitable access must be addressed to ensure that technological progress translates into improved care for all patients.

Within this rapidly evolving landscape, three major applications of AI in healthcare have become particularly prominent:

- 1) Opportunistic use of existing clinical data for large-scale, low-cost screening: contemporary healthcare ecosystems, increasingly reliant on the digitalization of patient information, generate immense volumes of imaging and structured clinical data. AI can effectively repurpose these datasets—originally collected for other clinical purposes—to screen asymptomatic populations and identify individuals at risk for high-burden diseases³.
- 2) AI systems that support and enhance clinical decision-making, especially through large-scale pre-trained or foundation models. These models learn general physiological representations from massive datasets, allowing them to perform multiple downstream tasks with minimal additional training⁴.
- 3) Mechanistic, multimodal, real-time simulations of biological systems—commonly known as digital twins. By integrating longitudinal signals, clinical variables, imaging, and sensor-derived physiological data, digital twins aim to reproduce the dynamic behavior of an individual patient's

cardiovascular system. These models hold promise for personalized medicine, early risk prediction, and virtual testing of therapeutic strategies before real-world intervention⁵.

This thesis, “*Applications of Artificial Intelligence in Cardiology: Methodological Advances and Three Clinical Use Cases*,” is structured around these three transformative domains. Each use case corresponds to one of the major applications of AI described above.

- 1) Use Case 1 explores the opportunistic use of deep learning models applied to chest X-rays imaging to predict cardiovascular outcomes through deep learning, demonstrating how routine imaging data can be leveraged for large-scale population screening.
- 2) Use Case 2 focuses on the development and evaluation of a self-supervised adaptable, robust, and interpretable ECG foundation model (HUBERT-ECG), illustrating the potential of pre-trained algorithms to enhance and generalize clinical decision-making.
- 3) Use Case 3 introduces the Dorian Gray Project, which uses AI to investigate the heart–brain axis and model the relationship between cardiovascular disease and cognitive decline, marking an early step toward constructing a cardiovascular digital twin.

Together, these studies highlight both the methodological innovation and the clinical relevance of AI-driven approaches in cardiology. They show how AI can improve prediction, support clinicians, and deepen our understanding of complex physiological systems—while also revealing the ethical, technical, and translational challenges that accompany this rapidly advancing field.

2. Use Case 1: Deep Learning for Predicting Cardiovascular Outcomes from Chest X-Ray Images

2.1 Background and Rationale

Atherosclerotic cardiovascular disease (ASCVD) remains the leading cause of mortality, premature death, and long-term morbidity worldwide, imposing a substantial and growing burden on healthcare systems and societies⁶. Within the spectrum of ASCVD, coronary artery disease (CAD) is responsible for the majority of fatal and non-fatal cardiovascular events, and early identification of individuals at increased risk remains one of the greatest challenges in preventive cardiology.

In the outpatient and diagnostic settings, the need for accurate, scalable, and cost-effective prediction strategies is particularly pressing. Traditional CAD prediction models rely on established clinical risk factors—age, sex, lipid levels, blood pressure, diabetes, and smoking history—and have been validated across selected cohorts^{7,8}. Despite their utility, these tools often perform suboptimally in heterogeneous real-world populations due to limited generalizability, evolving risk profiles, and the inherently multifactorial nature of CAD. Contemporary clinical guidelines therefore recommend estimating the pre-test probability (PTP) of obstructive CAD to guide decisions regarding non-invasive imaging or invasive coronary angiography⁹. However, existing PTP algorithms lack the granularity and precision required to reliably distinguish those who truly require further testing from those who do not, leading to both under- and over-utilization of diagnostic resources. As CAD prevalence continues to rise, there is a clear need for sensitive, reproducible, and broadly scalable methods capable of refining PTP estimation and improving clinical decision pathways.

Against this backdrop, artificial intelligence (AI) has emerged as a transformative technology capable of addressing several limitations inherent in traditional risk stratification. Advances in machine learning (ML) and deep learning (DL) have renewed interest in the diagnostic and prognostic potential of non-invasive imaging modalities, particularly those that are low-cost, widely accessible, and routinely integrated into clinical workflows¹⁰. AI-based image analysis can uncover subtle radiographic features and hidden phenotypes imperceptible to human observers, opening new avenues for individualized cardiovascular risk prediction and early disease detection¹¹. These innovations align with a broader shift toward opportunistic screening, wherein data acquired for one clinical purpose are repurposed to extract additional prognostic information without requiring new tests, additional radiation exposure, or increased healthcare costs.

Among all clinical imaging modalities, the chest radiograph (CXR) stands out as a uniquely powerful candidate for AI-enabled opportunistic cardiovascular screening. More than a century after its introduction, CXR remains the most frequently performed radiological examination worldwide¹². It

is inexpensive, fast, universally available, and routinely obtained across a wide range of clinical indications. Importantly, CXR is often performed early in the assessment of patients with suspected cardiopulmonary disease, making it one of the first imaging data points available to clinicians. Traditionally, CXRs are used to diagnose lung disease or identify gross cardiac abnormalities, yet the majority of examinations are reported as “normal,” and their potential to convey broader prognostic information has been largely underestimated.

In fact, a chest radiograph encodes a rich array of biological signals that extend far beyond its conventional interpretive boundaries. Subtle variations in cardiac silhouette geometry, pulmonary vascular markings, thoracic morphology, soft-tissue characteristics, skeletal integrity, and even signs of systemic comorbidities are captured within the radiographic projection¹³. Although clinicians review thousands of CXRs during their careers, fragmentary follow-up and the complexity of long-term disease trajectories prevent human readers from linking these early imaging features to future cardiovascular outcomes. As a result, much of the prognostic information contained in routine CXRs remains latent and underutilized.

AI provides a unique opportunity to unlock this hidden prognostic value. Deep learning models trained on large CXR datasets have demonstrated the ability to detect patterns associated with cardiopulmonary disease, vascular aging, metabolic dysfunction, and long-term mortality with a degree of consistency and precision unattainable through traditional human interpretation¹⁴. Several investigations have shown that AI applied to CXRs can predict all-cause mortality, incident cardiovascular events, or systemic conditions such as anemia and diabetes¹⁵. The first evidence that a DL model could derive long-term prognostic information from a single CXR emerged from large screening programs in oncology, where AI demonstrated the ability to forecast mortality beyond cancer-specific outcomes¹⁶. These findings suggest that CXRs contain underrecognized prognostic biomarkers that may be harnessed for cardiovascular risk prediction.

Applying AI to chest radiography for CAD assessment represents a promising frontier for cardiovascular medicine. Because CAD evolves gradually—often over decades—subclinical structural or physiological changes may be subtly encoded in the thoracic anatomy long before clinical symptoms arise. Deep learning algorithms may detect these early indicators, whether reflected in cardiac borders, vascular tracings, pulmonary patterns, tissue density gradients, or thoracic skeletal configurations, thereby enabling opportunistic risk stratification directly at the point of care. Crucially, such screening would require no additional resource use, as CXRs are already acquired in emergency departments, outpatient evaluations, and hospital admissions worldwide.

The rationale for the present work is therefore twofold.

First, the global burden of CAD demands innovative, scalable, and sustainable approaches capable of improving early diagnosis and risk prediction.

Second, chest X-ray imaging offers a unique combination of universality, affordability, and accessibility upon which AI-driven tools can be deployed at population scale. By leveraging DL models applied to routine CXRs, healthcare systems have the potential to enhance primary prevention strategies, reduce unnecessary invasive procedures, and improve the early detection of patients at heightened risk for adverse cardiovascular outcomes, including long-term mortality.

In this context, the research conducted in this use case comprises two complementary projects:

- 1) Development and clinical validation of a DL model designed to detect the presence of significant CAD from chest radiographs—demonstrating the feasibility of using CXR-based AI to support diagnostic triage and refine PTP estimation in patients referred for suspected CAD.
- 2) Evaluation of DL-based prediction of long-term mortality from plain CXR in a real-world cohort of patients referred for coronary angiography—extending the role of CXR beyond diagnosis to long-term prognostication, and illustrating how routine imaging can be repurposed for precision risk stratification.

Together, these studies contribute to the expanding field of opportunistic AI-driven cardiovascular screening, highlighting how existing imaging resources can be repurposed to support scalable, efficient, and personalized cardiovascular medicine.

Revolutionizing healthcare: artificial intelligence detection of coronary artery disease paves the way for future tools

Mauro Massussi^{a,b}, Marco Metra^a and Marianna Adamo^a

J Cardiovasc Med 2023, 24:467–468

^aDepartment of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia and ^bCardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, Brescia, Italy

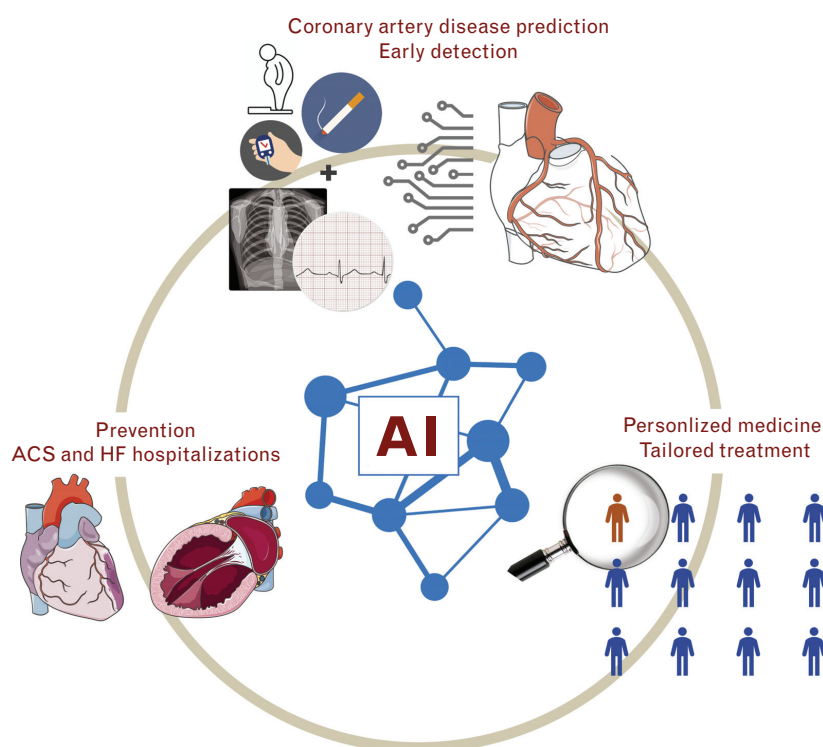
Correspondence to Mauro Massussi, MD, Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, 25123, Brescia, Italy
E-mail: mauro.massussi@gmail.com

Received 20 April 2023 Accepted 21 April 2023

Over the past few years, the integration of artificial intelligence has significantly improved early diagnosis and prediction of acute cardiac events for coronary artery

disease (CAD). CAD is one of the leading causes of death and morbidity worldwide, and acute coronary syndromes represent a significant manifestation of the disease that can have severe consequences for patients.^{1,2} By utilizing artificial intelligence, primary prevention strategies can be personalized to individual patients, resulting in a broader reach at a lower cost, and the ability to tackle unique challenges such as the low specificity of current diagnostic tools for patients admitted to emergency departments with suspected acute coronary syndromes (ACS).³

Fig. 1



The three key goals of artificial intelligence applied to the patients affected by coronary artery disease.

2.2 Deep Learning for Diagnostic Triage of Significant Coronary Artery Disease

Deep Learning to Detect Significant Coronary Artery Disease from Plain Chest Radiographs

Giuseppe D'Ancona, Mauro Massucci, Mattia Savardi, Alberto Signoroni, Lorenzo Di Bacco,
Davide Farina, Marco Metra, Roberto Maroldi, Claudio Muneretto, Hüseyin Ince, Davide
Costabile, Giuliano Chizzola, Salvatore Curello, Stefano Benussi

ABSTRACT

Aims: To train, test, and validate a deep learning (DL) algorithm for detecting the presence of significant coronary artery disease (CAD) based on chest radiographs.

Methods and Results: Data of patients undergoing chest radiography and coronary angiography were retrospectively analysed. A deep convolutional neural network (DCNN) was designed to detect significant CAD from the patient posteroanterior/anteroposterior chest radiograph. The DCNN was trained for binary classification of severe CAD absence/presence. Coronary angiography reports were used as ground truth. Information of 7728 patients referred for suspected angina was reviewed. Severe CAD was present in 4482 (58%) patients. Patients were randomly divided for training (70%; n=5454) and fine-tuning/testing (10%; n=773) of the algorithm. Internal clinical validation was performed with the remaining patients (20%; n=1501). At binary logistic regression, the DCNN prediction was the strongest independent determinant of severe CAD ($p < 0.0001$; OR: 1.040; CI: 1.032-1.048). Using an operating cut-point with high sensitivity, the DCNN had a sensitivity of 0.90 to detect significant CAD (AUC 0.73; 95% CI DeLong, 0.69-0.76). Adding to the AI chest radiograph interpretation angina status improved the prediction (AUC 0.77; 95% CI DeLong, 0.74-0.80).

Conclusion: Our DL algorithm can predict, with high sensitivity, the presence of severe CAD in patients referred for suspected angina. It could be used to pre-test significant CAD probability in outpatient clinics, emergency room settings, and CAD screening in more extensive settings. Further studies are required to externally validate the algorithm and develop a clinically applicable tool.

Introduction

Coronary artery disease (CAD) is the single leading cause of mortality, premature death, and morbidity worldwide (1; 2). CAD prediction models based on clinical risk factors have been validated against invasive coronary angiography and coronary computed tomography (CT) in limited cohorts. Moreover, derived risk scores are not specifically tailored to identify the presence of significant CAD that may require prompt treatment (3; 5). For this reason, patients are often referred to various diagnostic procedures, with suboptimal triaging and escalating degrees of invasiveness and costs for the healthcare budget. There is a general need to develop and apply more sensitive, practical, and sustainable CAD mass screening tools to rationalize and eventually expedite triage to more invasive coronary diagnostic investigations.

The recent introduction and popularization of artificial intelligence (AI) has resurged interests in the essential role that non-invasive imaging plays in diagnosing cardiovascular diseases (6, 7) and in supporting a more precise calculation of the individual's pre-test probability (PTP) of CAD. In this context, the predictive role of first-line diagnostic tools, such as chest radiographs, in patients with CAD may be underestimated and could benefit from AI-based applications. In the present manuscript, we summarize our experience with the training, testing, and clinically validating of a deep learning (DL) algorithm purposefully designed for detecting the presence of significant CAD based on chest radiographs.

Methods

Imaging and Data Curation

Data of patients referred to our institution for suspected angina and undergoing coronary angiography and chest radiography were retrospectively analysed. More than 85% of the chest radiographs and coronary angiography were performed within 18 days one from the other (**Supplementary figure 1**). Patients with a known history of CAD, previous history of percutaneous coronary revascularization, cardiac surgery, cardiac electronic devices implantation, and patients with low-quality images were excluded, to prevent biases in the algorithm training.

Only posteroanterior (PA) and anteroposterior (AP) chest radiographs were included in the study and later-lateral radiographs were excluded because they were not consistently present in all patients. All images (computed radiography and digital radiography) were stored in the Institutional Picture Archiving and Communication System (PACS).

Coronary angiography reports were produced by experienced interventional cardiologists, prospectively categorized, and digitalized in an electronic database at the time of investigation

together with information concerning the patient's age, gender, body mass index (BMI), angina status, and Diamond-Forrester score (age, gender, angina status).

Algorithm Settings

A deep convolutional neural network (DCNN) was trained to detect the presence of significant CAD from the patient chest radiograph. The ground truth reference of CAD was structured using the coronary angiography reports previously produced. Two groups were identified with either no significant CAD (complete absence of CAD or presence of $< 70\%$ coronary stenosis) or with significant CAD (at least one coronary stenosis $\geq 70\%$ in at least one of the main coronary vessels, i.e., left main, left anterior descending, left circumflex, and right coronary artery). The system was trained for binary classification of absence or presence of severe CAD (0, 1). Patients were randomly divided for training (70%) and DCNN model tuning (10%). Internal validation was performed with the remaining patients (20%).

Statistical Analysis

Statistical analysis was performed using the Python scikit-learn and scipy libraries. Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) were calculated to evaluate the binary classifier performance of the DCNN at different thresholds and achieve the highest sensitivity or the maximum sum of sensitivity and specificity. AUCs were compared using the DeLong parametric method, and 95% confidence intervals (CI) were produced using two-sided CIs for proportions. Multivariable analysis using binary logistic regression was performed on the internal validation set to identify independent correlation among ML prediction, BMI, age, gender, Diamond-Forrester score (independent variables), and the presence of significant CAD (dependent variable). The statistical significance threshold was $p\text{-value} < 0.05$. Finally, GradCAM heat maps were generated to define the chest radiograph features supporting the different DCNN decisions (8).

Results

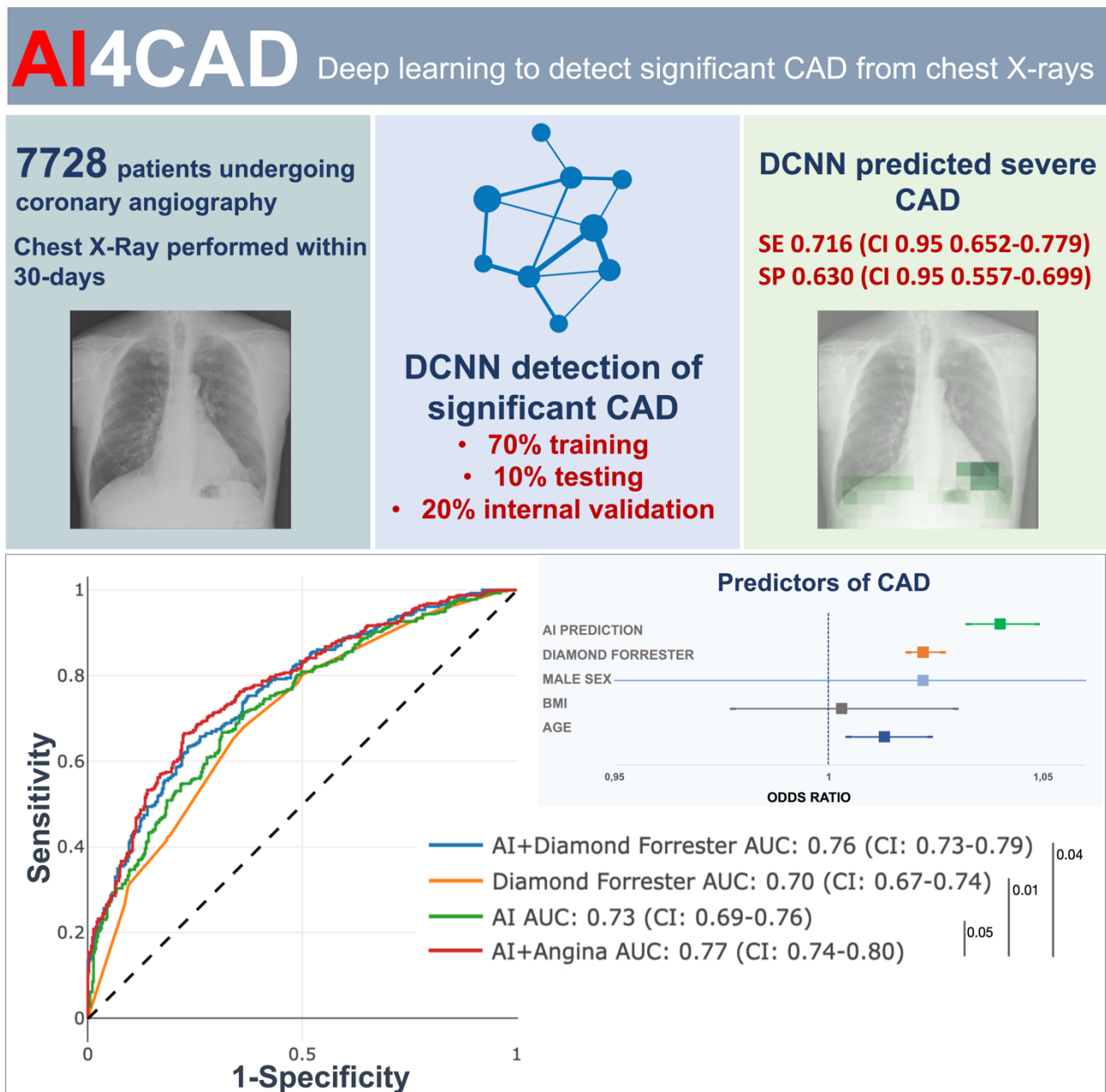


Figure 1. Graphical Abstract

Patients. Information of 10097 patients referred to our institution for coronary angiography was reviewed. After selection according to the criteria mentioned above, we included 7728 patients (Figure 1). Severe CAD was present in 4482 (58%) patients (2% left main, 47% one vessel, 31% two vessels, and 20% 3 vessels disease). **Supplementary Table 1** summarizes demographic/clinical information in the cohort and severity/distribution of CAD.

DCNN. Patients were randomly divided for training (70%; n=5454) and fine-tuning/testing (10%; n=773) of the algorithm. Internal validation was performed with the remaining patients (20%; n=1501). The DCNN had a processing time of about 20 sec to analyse the whole internal validation set on a workstation with an Nvidia Titan V Graphics Processing Unit (GPU) with 12GB of memory.

Clinical characteristics of patients who tested negative compared to those who tested positive to the DCNN-based chest radiography prediction of severe CAD are reported in **Table 1**. Using an operating cut-point with high sensitivity, the algorithm had sensitivity of 0.90 and specificity of 0.31 to detect significant CAD in the internal validation group (AUC 0.73; 95% CI DeLong, 0.69-0.76). The operating point leading to the maximum sensitivity-specificity sum produced a sensitivity of 0.71 [CI 0.66-0.77] and a specificity of 0.63 [CI 0.56-0.6] (**table 2**). Adding to the AI-based chest radiograph interpretation angina status improved the prediction (AUC 0.77; 95% CI DeLong, 0.74-0.80). Finally, the Diamond-Forrester prediction achieved an AUC of 0.70 (95% AUC CI: 0.66-0.73) with the point of maximum sensitivity-specificity sum having a sensitivity of 0.68 (CI 0.61-0.74) and specificity of 0.63 (CI 0.56-0.70) (**Figure 2**).

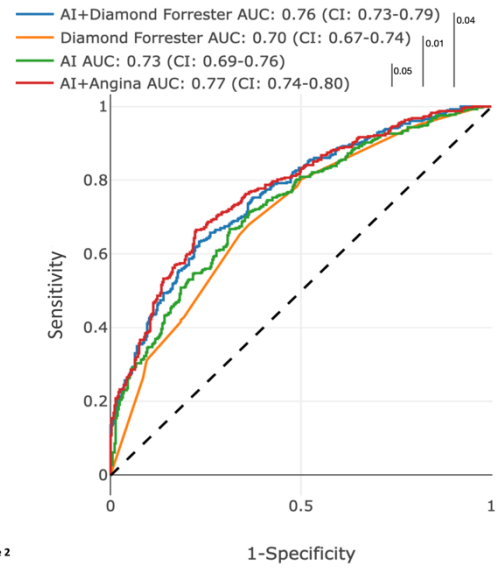


Figure 2. Receiver operating characteristic (ROC) curves for the binary classification of the presence significant coronary artery disease (CAD) demonstrating area under the curve (AUCs) of 0.77 for AI + angina Type; 0.76 for AI + Diamond Forrester; 0.73 for AI

To understand which kind of information flows through the DL model, we removed the last layer in charge of identifying the CAD risk, and we used the obtained intermediate representation to train a logistic regression model that tries to estimate age, gender, and BMI. The predicted age has a correlation of 0.38 ($p < 0.001$) with respect to the real one, while the BMI has a correlation of 0.54 ($p < 0.001$). Finally, the gender prediction reaches 86% accuracy (specificity: 73%; sensitivity: 92%). The C-statistic of the multivariable model to predict presence of severe CAD was 0.75 (95% CI: 0.70-0.79). At binary logistic regression the DCNN prediction was the strongest independent determinant of severe CAD ($p < 0.001$; OR: 1.040; CI: 1.032-1.048) (**table 3**). Starting from the chest radiograph the DCNN could differentiate patients according to presence and extension of severe CAD. **Figure 3** shows the comparison, within the validation group, of the scores attributed from the DCNN according to the distribution of CAD severity and its extension (moderate noncritical

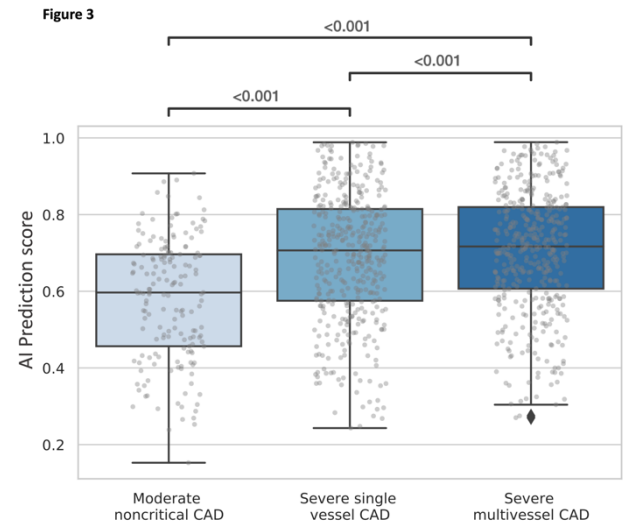


Figure 3. Comparison, within the validation group, of the scores attributed from the DCNN according to the distribution of coronary artery disease (CAD) severity and its extension (moderate noncritical CAD, severe single vessel CAD, severe multivessel CAD).

CAD, severe single vessel CAD, severe multivessel CAD).

Heat Maps. Attention maps were created considering the DCNN with highest performance. Heat maps primarily localized to the cardiac silhouette, left ventricular apex, pulmonary bases, pulmonary parenchyma, costophrenic sinuses, pulmonary hila, thoracic aorta, supra-aortic vessels, and clavicle region (**Figure 4, panel A-F**).

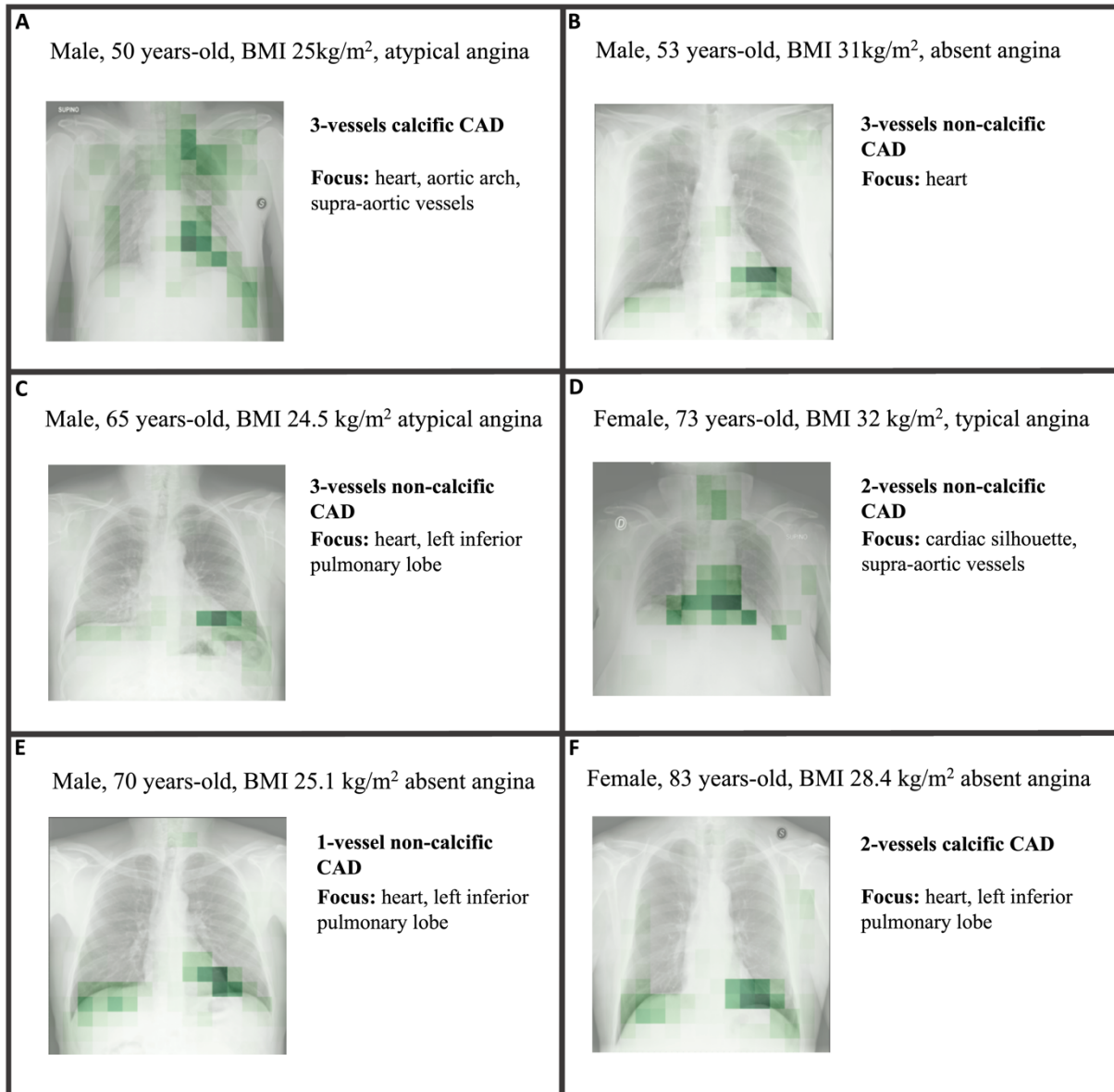


Figure 4. (A-F) Heat-maps of 6 patients affected by severe CAD. Areas suggestive of CAD presence are highlighted in degrees of green tonality

Discussion

Study findings can be summarized as follows: 1) Our DL algorithm can predict, from a one projection chest radiograph, the presence of severe CAD; 2) Results are confirmed with invasive coronary angiography and are internally validated; 3) At internal validation, the DCNN could differentiate severity and extension of CAD; 4) The algorithm also has a good prediction of accepted CAD risk

scores, such as age, gender, and BMI; 4) Although the accuracy of the DCNN prediction is superior to that achieved by accepted risk scores such as the Diamond-Forrester, adding the angina type to the model significantly improves the severe CAD prediction of the DCNN.

CAD is the leading global health burden, and, for this reason, sensitive, practical, and sustainable CAD screening tools are required to rationalize and guide its diagnosis and eventual treatment. Guidelines suggest that an individual's PTP of CAD should be calculated in those presenting with chest pain (9,10). Different risk scores have been developed to categorize risk based on the clinical history and prompt triage to more costly, invasive, and less ubiquitous diagnostic modalities (3, 4). Although non-invasive first-line imaging tools could play an essential role in supporting CAD diagnosis in patients referred for suspected angina, their potential remains unexploited. Chest radiograph, one of the first imaging investigations performed in patients referred for chest pain and suspected angina, is, for example, considered neither sensitive nor specific for the actual diagnosis of CAD and, for this reason, it is hardly kept into consideration when stratifying patients at risk of CAD. Concurrently, chest radiographs are ubiquitous, have fast acquisition, and carry a plethora of information concerning the patient's overall and specific health status¹⁷. Previous literature has shown that specific areas in the chest radiograph may present direct and indirect signs of cardiovascular pathology, including significant CAD.

Watanabe *et al.* have recently demonstrated that aortic arch calcification with pericardial fat mass detected on a single chest radiograph image is closely associated with predictive variables of cardiovascular diseases, such as the carotid intima-media thickness, cardio-ankle vascular index, the Suita score, and other cardiovascular risk points (11). Moreover, there is growing evidence that the burden of calcification at the aortic valve and mitral ring level, which can be easily detected with the chest radiograph, is directly related to the presence of CAD and the extent and composition of coronary arteries plaque (12, 13). Furthermore, chest radiographs contain essential information about the patient body habitus, the pulmonary condition (including lung congestion, interstitial lung disease, chronic obstructive pulmonary disease), and the bone mineral density (at the clavicle and vertebral bodies level), which are all related to the presence of CAD (14, 17). Finally, even simple geometrical measurements derived from the chest radiograph, such as, for example, cardiothoracic ratio, can independently predict cardiovascular mortality in patients undergoing coronary angiography (18).

The introduction of AI-based interpretation of medical images has demonstrated the potential to detect markers of pathology that remain overlooked by human observers and improve image interpretation across a significant breadth of clinical practice (19). AI offers the unprecedented possibility to discover even new and previously unexplored hidden features of disease within clinical data. DL, a type of AI, can unleash complex relationships in large data sets, like those present within

diagnostic images and make predictions that would not be otherwise supported by the most used models of cardiovascular risk stratification. AI-based interpretation of ubiquitous and readily available imaging may lead to a sensitive prediction of CAD and, in this way, open future possibilities for opportunistic screening in larger populations.

In previous studies AI has been used to automatically compute coronary arteries calcium (CAC) scores from ECG-gated cardiac CT and non-contrast CT images (20, 21). Although often used as a first-line screening tool in patients with suspected angina, CT scans have higher costs, radiation burden, and are less available than chest radiographs. In a recent study, Kamel et al. have shown, in a small cohort of 1010 patients, that DCNNs can predict the likelihood of CACs on chest radiographs (22), using cardiac CT as ground truth and achieving a sensitivity of 91% and specificity of 29% in a sensitivity-optimized algorithm. Although the authors did not focus on predicting CAD severity but rather on the presence/absence of CAC, their findings are encouraging and confirm the potential of AI-based chest radiograph interpretation for the definition of CAD presence.

Our study focused on the specific prediction of severe CAD, defined as at least one coronary stenosis $\geq 70\%$ in at least one of the main coronary vessels. Our model, differently from the one proposed by Kamel et al. (22), was not trained to derive the diagnosis of CAD based on the simple recognition of coronary calcifications at chest radiograph. The presence of CACs is not necessarily related to significant coronary lesions. Furthermore, the quality of chest radiographs can be suboptimal to detect coronary vessels calcifications, even when read by an intelligent system whose image resolution and recognition are superior to those of the human eye.

As ground truth to train and test our DCNN, we used reports from the gold-standard method for CAD severity definition, i.e., invasive coronary angiography, that gives information beyond simple calcium detection. In our semi-supervised system, we did not limit the DCNN analysis on specific areas of the chest radiograph, like for example the cardiac region. Consequently, and as emerging from the attention maps produced from the chest radiographs, the areas supporting a positive attribution of severe CAD were not necessarily defined by the presence of peculiar patterns at the level of the cardiac silhouette. Although to make light on the "black box" of the ML algorithms decision-making process is not an easy task, we have seen that the DCNN has also identified chest radiographs patterns outside of the isolated cardiac region and has used them to rule out the presence of severe CAD.

The trained DCNNs can predict commonly used CAD risks stratification factors, such as age, BMI, and gender, and possibly consider them when ruling in the presence of severe CAD (23, 24). In logistic regression, the DCNN predicts severe CAD independently by the risk factors, suggesting that the system provides supplementary information in the risk assessment model. This finding is

confirmed because the DCNN has overcome the prediction of already accepted and used screening tools based on clinical factors, like the Diamond-Forrester score.

As expected, the algorithm cannot easily predict clinical variables such as angina status from the chest radiograph. This information is crucial in identifying patients at risk of having severe CAD. Adding the information concerning the angina status to the DCNN allowed us to improve the accuracy in predicting severe CAD (AUC from 0.73 to 0.77), with values of sensitivity around 90%. Finally, we must emphasize performing an adequate image mining to prevent training biases. Thanks to the timely screening that allowed us to exclude patients with previous cardiac interventions, including cardiac surgery, PCI, and cardiac electronic devices implantation, we identified no secondary confounding factors in the radiographs that may have influenced and biased the DCNN decision process and identification of CAD patients. In our training and testing phases, we have included chest radiographs performed in bed and the standing position, in emergent and elective settings, and images produced by nine different portable and stationary radiographic machines available in our Institution. This research setting reflects the everyday clinical practice, reduces the risks of distributional shift, and enhances the chances of success for a future external validation phase.

Conclusion

Our DL algorithm trained on coronary angiograms and plain chest radiographs can predict, with high sensitivity, the presence of severe CAD in patients referred for invasive coronary investigations. The algorithm optimized for sensitivity has shown a 10% false-negative rate and could be used as a tool for a more rational and prompt triaging to diagnosis/treatment of patients with suspected angina, using escalating degrees of invasiveness. These results were achieved within a study cohort with a high prevalence (58%) of severe CAD, where a confirmatory invasive angiography had been required to plan treatment.

Chest radiographs are ubiquitous, and the algorithm could be used with the angina status information for pre-test significant CAD probability assessment in outpatient clinics and emergency room settings. Further studies are required to externally validate the algorithm, evaluate its performance in broader cohorts with a lower prevalence of CAD, optimize its ability to distinguish among different degrees of CAD extension, and develop a practical tool for everyday clinical applicability that could also provide information about patient's longevity and prognosis¹⁸.

References

1. Li, Z. et al. Global, Regional, and National Death, and Disability-Adjusted Life-Years (DALYs) for Cardiovascular Disease in 2017 and Trends and Risk Analysis From 1990 to 2017 Using the Global Burden of Disease Study and Implications for Prevention. *Frontiers in public health* 9, 559751 (2021).
2. Timmis, A. et al. European Society of Cardiology: cardiovascular disease statistics 2021. *European heart journal* 43, 716–799 (2022).
3. Baskaran, L. et al. A Comparison of the Updated Diamond-Forrester, CAD Consortium, and CONFIRM History-Based Risk Scores for Predicting Obstructive Coronary Artery Disease in Patients With Stable Chest Pain: The SCOT-HEART Coronary CTA Cohort. *JACC. Cardiovascular imaging* 12, 1392–1400 (2019).
4. Genders, T. S. S. et al. A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. *European heart journal* 32, 1316–1330 (2011).
5. Almeida, J. et al. Comparison of Coronary Artery Disease Consortium 1 and 2 Scores and Duke Clinical Score to Predict Obstructive Coronary Disease by Invasive Coronary Angiography. *Clinical cardiology* 39, 223–228 (2016).
6. Dey, D. et al. Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review. *Journal of the American College of Cardiology* 73, 1317–1335 (2019).
7. Bader, A. S., Rubinowitz, A. N., Gange, C. P. J., Bader, E. M. & Cortopassi, I. O. Imaging in the Evaluation of Chest Pain in the Primary Care Setting, Part 1: Cardiovascular Etiologies. *The American journal of medicine* 133, 1033–1038 (2020).
8. Selvaraju, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. in 2017 IEEE International Conference on Computer Vision (ICCV) 618–626 (2017). doi:10.1109/ICCV.2017.74.
9. Fihn, S. D. et al. 2014 ACC/AHA/AATS/PCNA/SCAI/STS focused update of the guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, . *Journal of the American College of Cardiology* 64, 1929–1949 (2014).
10. Knuuti, J. et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. *European heart journal* 41, 407–477 (2020).
11. Watanabe, K. et al. Aortic arch calcification with pericardial fat mass detected on a single chest X-ray image is closely associated with the predictive variables of future cardiovascular disease. *Heart and vessels* 37, 654–664 (2022).
12. Mahabadi, A. A. et al. Association of aortic valve calcification to the presence, extent, and composition of coronary artery plaque burden: from the Rule Out Myocardial Infarction using Computer Assisted Tomography (ROMICAT) trial. *American heart journal* 158, 562–568 (2009).
13. Schneider, M. et al. Predicting the presence of coronary artery disease by transesophageal echocardiography. *Wiener klinische Wochenschrift* 132, 708–715 (2020).
14. Clarson, L. E. et al. Interstitial lung disease is a risk factor for ischaemic heart disease and myocardial infarction. *Heart (British Cardiac Society)* 106, 916–922 (2020).
15. Boschetto, P., Beghé, B., Fabbri, L. M. & Ceconi, C. Link between chronic obstructive pulmonary disease and coronary artery disease: implication for clinical practice. *Respirology (Carlton, Vic.)* 17, 422–431 (2012).
16. Zhang, Y., He, B., Wang, H., Shi, J. & Liang, H. Associations between bone mineral density and coronary artery disease: a meta-analysis of cross-sectional studies. *Archives of osteoporosis* 15, 24 (2020).
17. Kumar, D. A. & Anburajan, M. The role of hip and chest radiographs in osteoporotic evaluation among south Indian women population: a comparative scenario with DXA. *Journal of endocrinological investigation* 37, 429–440 (2014).

18. Zaman, M. J. S. et al. Cardiothoracic ratio within the ‘normal’ range independently predicts mortality in patients undergoing coronary angiography. *Heart (British Cardiac Society)* 93, 491–494 (2007).
19. Seah, J. C. Y. et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet. Digital health* 3, e496–e506 (2021).
20. Wolterink, J. M., Leiner, T., Takx, R. A. P., Viergever, M. A. & Isgum, I. Automatic Coronary Calcium Scoring in Non-Contrast-Enhanced ECG-Triggered Cardiac CT With Ambiguity Detection. *IEEE transactions on medical imaging* 34, 1867–1878 (2015).
21. Shadmi, R., Mazo, V., Bregman-Amitai, O. & Elnekave, E. Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) 24–28 (2018).
22. Kamel, P. I., Yi, P. H., Sair, H. I. & Lin, C. T. Prediction of Coronary Artery Calcium and Cardiovascular Risk on Chest Radiographs Using Deep Learning. *Radiology. Cardiothoracic imaging* 3, e200486 (2021).
23. Raghu, V. K., Weiss, J., Hoffmann, U., Aerts, H. J. W. L. & Lu, M. T. Deep Learning to Estimate Biological Age From Chest Radiographs. *JACC. Cardiovascular imaging* 14, 2226–2236 (2021).
24. Lu, M. T. et al. Deep Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA network open* 2, e197416 (2019).

Supplementary figure 1. Time distribution of the Chest Radiographs (CXR) with respect to coronary angiography. Negative values represent a CXR taken before the coronary angiography.

Supplementary figure 2. Chest Radiographs (CXR) grouped according to different manufacturer.

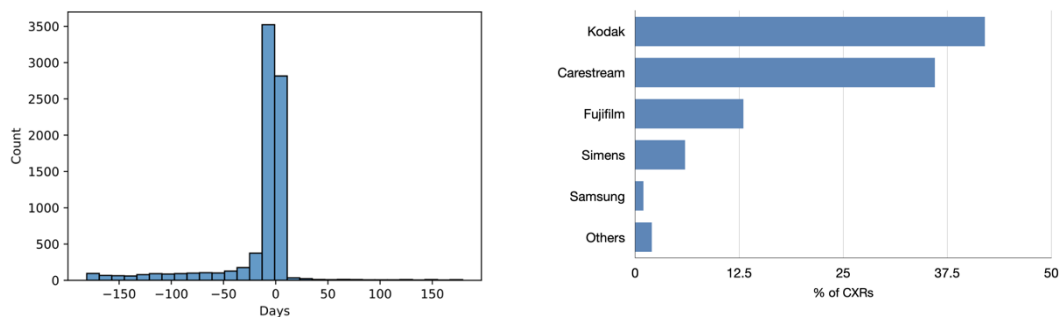


Table 1. Demographic/clinical information in patients who tested negative and positive on the DCNN-based chest radiography prediction of severe CAD. Continuous variables presented with mean and standard deviation.

	Algorithm tested negative	Algorithm tested positive	<i>p-value</i>
Age	75±13	74±12	0.172
BMI	25±4	27±4	<0.001
Sex (male)	37%	77%	<0.001
Absent Angina	43%	32%	<0.001
Atypical Angina	33%	31%	<0.001
Typical Angina	24%	37%	<0.001
Diamond Forrester score	47.2±29.4	61.3±29.2	<0.001

BMI=body mass index

Table 2. DCNN performance for prediction of significant CAD on chest radiographs compared to Diamond-Forrester performance.

	Sensitivity	Specificity
Diamond Forrester	0.94 (CI 0.95 0.91-0.97) 0.68 (CI 0.95 0.62-0.75)	0.22 (CI 0.95 0.16-0.28) 0.64 (CI 0.95 0.57-0.71)
AI	0.90 (CI 0.95 0.86-0.94) 0.72 (CI 0.95 0.65-0.78)	0.32 (CI 0.95 0.25-0.38) 0.63 (CI 0.95 0.56-0.70)
AI + Diamond Forrester	0.90 (CI 0.95 0.86-0.94) 0.64 (CI 0.95 0.57-0.70)	0.35 (CI 0.95 0.28-0.41) 0.76 (CI 0.95 0.70-0.82)
AI + Angina Type	0.90 (CI 0.95 0.86-0.94) 0.67 (CI 0.95 0.60-0.73)	0.36 (CI 0.95 0.30-0.44) 0.77 (CI 0.95 0.71-0.84)

AI=artificial intelligence

Table 3. Predictors of CAD at logistic regression analysis

	OR (CI 95%)	<i>p-value</i>
Age	1.013 (0.978-0.991)	0.006
BMI	1.003 (0.977-1.029)	0.842
Sex (male)	1.022 (0.783-1.333)	0.873
Diamond Forrester score	1.022 (1.018-1.026)	<0.001
AI prediction	1.040 (1.032-1.048)	<0.001

BMI=body mass index; AI=artificial intelligence

Supplementary table 1. Baseline characteristics of the examined population

	Study population (N=7728)
Sex (male)	5429 (71%)
Age (years)	74 ± 12
BMI	26 ± 4
Coronary angiography	
Absent CAD	2911
Moderate CAD	726
Left Main CAD	68
1 vessel CAD	1996
2 vessels CAD	1165
3 vessels CAD	862

BMI=body mass index; CAD=coronary artery disease



Contents lists available at ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard

Deep learning to detect significant coronary artery disease from plain chest radiographs AI4CAD

Giuseppe D'Ancona^{a,*}, Mauro Massucci^b, Mattia Savardi^c, Alberto Signoroni^c, Lorenzo Di Bacco^d, Davide Farina^e, Marco Metra^b, Roberto Maroldi^e, Claudio Muneretto^d, Hüseyin Ince^a, Davide Costabile^f, Monica Murero^g, Giuliano Chizzola^b, Salvatore Curello^b, Stefano Benussi^d

^a Department of Cardiology and Cardiovascular Clinical Research Unit, Vivantes Klinikum Urban and Neukölln, Berlin, Germany

^b Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili and Department Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, Italy

^c Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, and Department of Information Engineering, University of Brescia, Brescia, Italy

^d Department of Cardiac Surgery, Spedali Civili Brescia and University of Brescia, Brescia, Italy

^e Radiology 2, ASST Spedali Civili and Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, Italy

^f Department of Information Technology Spedali Civili Brescia, Brescia, Italy

^g AI4 Life and Society International Institute, Federico II University, Naples, Italy

ARTICLE INFO

Keywords:

Coronary Artery Disease
Chest Radiograph
Artificial Intelligence
Deep Learning

ABSTRACT

Background: The predictive role of chest radiographs in patients with suspected coronary artery disease (CAD) is underestimated and may benefit from artificial intelligence (AI) applications.

Objectives: To train, test, and validate a deep learning (DL) solution for detecting significant CAD based on chest radiographs.

Methods: Data of patients referred for angina and undergoing chest radiography and coronary angiography were analysed retrospectively. A deep convolutional neural network (DCNN) was designed to detect significant CAD from posteroanterior/anteroposterior chest radiographs. The DCNN was trained for severe CAD binary classification (absence/presence). Coronary angiography reports were the ground truth. Stenosis severity of $\geq 70\%$ for non-left main vessels and $\geq 50\%$ for left main defined severe CAD.

Results: Information of 7728 patients was reviewed. Severe CAD was present in 4091 (53%). Patients were randomly divided for algorithm training (70%; $n = 5454$) and fine-tuning/model validation (10%; $n = 773$). Internal clinical validation (model testing) was performed with the remaining patients (20%; $n = 1501$). At binary logistic regression, DCNN prediction was the strongest severe CAD predictor ($p < 0.0001$; OR: 1.040; CI: 1.032–1.048). Using a high sensitivity operating cut-point, the DCNN had a sensitivity of 0.90 to detect significant CAD (specificity 0.31; AUC 0.73; 95% CI DeLong, 0.69–0.76). Adding to the AI chest radiograph interpretation angina status improved the prediction (AUC 0.77; 95% CI DeLong, 0.74–0.80).

Conclusion: AI-read chest radiographs could be used to pre-test significant CAD probability in patients referred for suspected angina. Further studies are required to externally validate our algorithm, develop a clinically applicable tool, and support CAD screening in broader settings.

1. Introduction

Coronary artery disease (CAD) is the single leading cause of mortality, premature death, and morbidity worldwide [1,2]. In limited

cohorts, CAD prediction models based on clinical risk factors have been validated against invasive coronary angiography and coronary computed tomography (CT) [3–5]. In clinical practice, specific risk factors should be evaluated to define obstructive CAD pre-test

* Corresponding author at: Cardiovascular Clinical Research Unit at Department of Cardiology, Vivantes Klinikum Am Urban Dieffenbachstraße 1, 10967 Berlin, Germany.

E-mail address: rgea@hotmail.com (G. D'Ancona).

<https://doi.org/10.1016/j.ijcard.2022.10.154>

Received 20 August 2022; Received in revised form 18 October 2022; Accepted 24 October 2022

Available online 5 November 2022

0167-5273/© 2022 Elsevier B.V. All rights reserved.

2.3 Long-Term Mortality Prediction using Deep Learning on Chest X-Rays

Plain Chest Radiographs Predicts Long-Term Mortality: Artificial Intelligence Approach in Patients Referred for Suspected Coronary Artery Disease

Giuseppe D'ancona, Mattia Savardi, Mauro Massucci, Viktor Van Der Valk, Roderick W.C.

Scherptong, Alberto Signoroni, Davide Farina, Monica Murero, Hüseyin Ince, Stefano Benussi, Salvatore Curello, Fatih Arslan

ABSTRACT

Background: Deep learning (DL) based interpretation of chest X-ray (CXR) has been shown to predict follow-up mortality of patients screened for neoplastic disease.

Purpose: We summarize our experience with a DL model to predict, from a single projection CXR, long-term mortality of patients referred for angina and coronary angiography.

Methods: Demographic, clinical, and follow-up data of patients referred to our institution (Italy) were analyzed retrospectively. A deep convolutional neural network (DCNN) was designed to predict follow-up mortality from CXRs. External validation was performed on patients referred to a medical institution in the Netherlands.

Results: 6031 were randomly divided for model training (70%; n=4259) and model fine-tuning/model validation (10%; n=602). Internal clinical validation was performed with the remaining patients (20%; n=1170). Patients were stratified according to DL-CXR risk score quartiles. Median follow-up was 6.1 years (IQR 3.3 – 8.7). The DL-CXR risk score was the strongest predictor of follow-up mortality ($p<0.005$; OR: 3.16; CI: 2.47-3.85) and had an AUC of 0.793 (95% AUC CI: 0.759-0.827, Sens. 78%, Spec. 0.68%). The prediction model improved when age and coronary artery disease (CAD) status were included (AUC 0.81; CI: 0.775-0.843). Estimated mortality increased significantly with DL-CXR risk score (Low-risk 5%, Moderate 17%, High 29%, Very High 46%; $p<0.0001$). Results on external validation confirmed the DL-CXR-risk score prediction performance (AUC 0.712 CI: 0.495-0.928; Sens 0.838; Spec 0.338).

Conclusion: In patients with suspected angina referred for coronary angiography, the DL CXR-risk score could stratify long-term mortality risk better than age, CAD status, sex, and BMI.

Introduction

The chest radiograph (CXR) is one of the first imaging investigations performed on patients referred for suspicion of cardiac conditions and possibly the most common imaging test performed in the medical field [1, 2]. Most CXRs are reported as non-pathologic and are used to rule out a specific clinical suspicion mostly related to the lungs. However, CXRs carry a plethora of different information, often completely undiscovered and unreported, concerning the patient's overall and specific health status [3]. When identified, this information could help make the diagnosis at the time of referral, support risk stratification of future morbidity and mortality, and guide a chain of decisions concerning lifestyle changes, screening planning, and prevention strategies. Clinicians and radiologists evaluate many CXRs during their medical career but are rarely able to follow patients in their long-term clinical developments, and, for this reason, they cannot connect and articulate the multifaceted radiographic findings at time zero with events occurring at long-term follow-up.

The recent introduction and popularization of artificial intelligence (AI) have resurged interest in the essential role that non-invasive ubiquitous and affordable imaging modalities such as CXRs may play in diagnosing pathologies and eventually helping define outcomes at follow-up. Deep learning (DL) has demonstrated the potential to detect markers that remain overlooked by human observers and could improve image interpretation across many clinical practices [4]. Apart from the emerging potential that AI is demonstrating for the enhancement and support of diagnosis at the time of patient referral, the use of AI to define a patient's prognosis starting from simple medical images, offers even more fascinating, and challenging, perspectives. The hypothesis that a DL model can produce long-term prognostic information from CXR has already been confirmed within a lung and prostate screening program [5]. In this manuscript we summarize our experience with the training, testing, and clinically validating of a DL algorithm purposefully designed for predicting, from a single projection CXR, long-term mortality of patients referred for chest pain and suspected coronary artery disease (CAD).

Methods

Imaging and Data Curation. The patients' cohort included in the present study has been already used by our group to test, train, and validate a deep convolutional neural network (DCNN) solution for detecting significant CAD based on chest radiographs [6].

Data of patients referred to our institution (Spedali Civili di Brescia, Brescia, Italy) for suspected angina were retrospectively analysed. Radiographs were performed using nine different radiographic machines. Exclusion criteria to prevent algorithm training biases included: previous history of

percutaneous coronary revascularization, cardiac surgery, cardiac electronic devices implantation, and low-quality acquisition.

Only posteroanterior (PA) and anteroposterior (AP) CXRs were included in the study, and lateral radiographs were excluded because they were not consistently present in all patients. All images were stored in the institutional Picture Archiving and Communication System (PACS).

Coronary angiography reports, produced by interventional cardiologists, were prospectively collected, and categorized in an electronic database at the time of investigation. Severe CAD was defined for patients with at least one visually estimated coronary diameter stenosis with severity of $\geq 70\%$ for non-left main disease and $\geq 50\%$ for left main disease [6]. All other patients were classified as having no significant/severe CAD. We only used physiological assessment with fractional flow reserve (FFR) and instantaneous wave-free ratio (iFR) in coronary lesions of uncertain significance. Results were incorporated to draw the final diagnosis.

Demographic and clinical information concerning the patient's age, biological sex, body mass index (BMI), angina status at time of referral were also collected [7].

For every patient included in the cohort, demographic information contained in the hospital electronic medical records (EMR) was matched with the Italian National Census data, to document follow-up status (dead or alive) in August 2022.

Patients had signed informed consent to use relevant clinical information for scientific purposes. All data have been anonymized, and the study protocol has been approved by the local scientific and ethical committee.

Algorithm Settings. A DCNN was trained to detect long-term clinical- outcome from the patient CXR at time of referral for coronary angiogram. The ground truth reference of follow-up status was structured using official data from the Italian census. Based on the follow-up status, we identified two groups, dead or alive at long-term follow-up.

We trained the system for binary classification (0, Alive; 1 Dead). Patients were randomly divided for training (70%) and DCNN model tuning (10%). We performed internal validation (model testing) with the remaining patients (20%).

Given the workload and the numerosity of training samples, we selected a DCNN pre-trained on PadChest [8], NIH ChestX-ray8 [9], CheXpert [10], and MIMIC [11] datasets. In particular, after an extensive model search, we found ResNet50 to be the most suited [12]. We exploited TorchXRayVision backed up with PyTorch on Python 3.7 for the model training and validation [13]. The performed hyperparameters search to minimize validation loss was carried out using a binary cross-entropy with 100 maximum epochs and early stopping to avoid overfitting. To improve

robustness, data augmentation was performed by applying random geometric, brightness, and contrast transformations. All the images underwent a standard normalization between 0 and 1.

External Validation. After internal validation of the DCNN, we performed external validation using a sample of 193 patients referred for chest pain to a different medical institution located in north-western Europe, the Leiden University Medical Center, (Leiden, the Netherlands). Inclusion criteria in the external validation group were the same as those used for the inclusion in the original dataset.

Statistical Analysis. Statistical analysis was performed using the Python scikit-learn and scipy libraries. Sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC) were calculated to evaluate the binary classifier performance of the DCNN at different thresholds and achieve the highest sensitivity or the maximum sum of sensitivity and specificity. AUCs were compared using the DeLong parametric method, and 95% confidence intervals (CI) were produced using two-sided CIs for proportions. The association between the DL-derived chest radiograph-risk score and follow-up all-cause mortality (primary outcome) was tested using Cox proportional hazards regression models and Kaplan-Meier (KM) statistics. The ROC AUCs were used to further assess follow-up mortality discrimination of the AI derived radiography-risk score and compare it to that of already known risk-factors (age, sex, body mass index, presence and extension of CAD, and angina status). The statistical significance threshold was a p-value < 0.05. Finally, GradCAM heat maps were generated to define the chest radiograph features supporting the different DCNN decisions [14].

Results

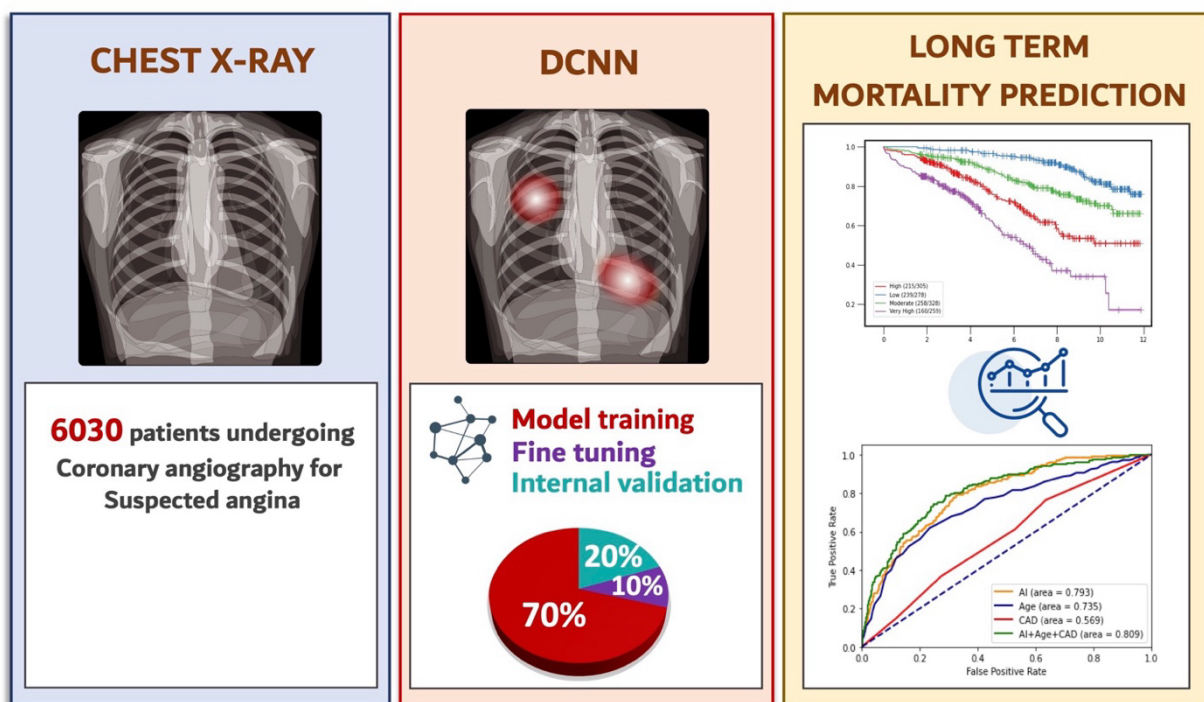


Figure 1. Graphical abstract

Patients: A total of 6031 patients was randomly divided for model training (70%; n=4259) and model fine-tuning/model validation (10%; n=602). Internal clinical validation (model testing) was performed with the remaining patients (20%; n=1170). Demographic and clinical information, including presence and eventual degree of CAD are presented in **Table 1**.

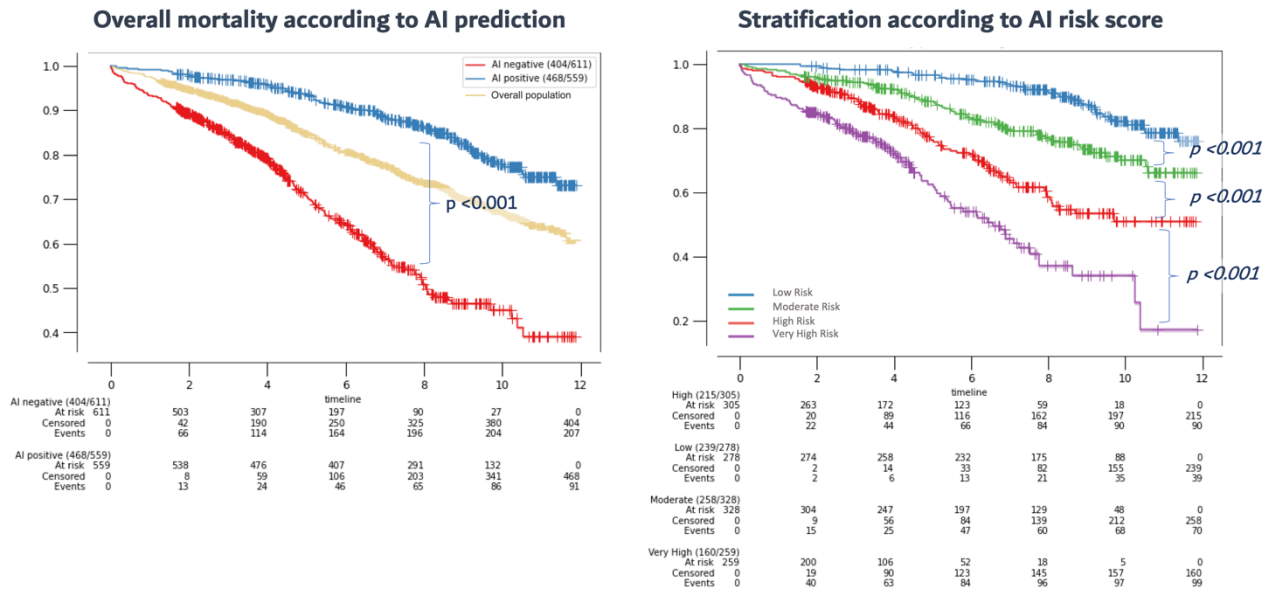


Figure 2. Kaplan-Meier curves for all-cause mortality in the overall population and two groups stratified by AI prediction. (A) Follow-up curve. (B) Quartile analysis of AI prediction

DCNN: Patients were stratified according to DL-based interpretation quartiles (DL-CXR risk score). Median follow-up was 6.1 years (IQR 3.3 – 8.7) with a KM overall estimated mortality of 21.9% (CI 19.2%-24.5%) (**Figure 2, panel A**). Estimated mortality increased significantly according to the DL-CXR risk score (Low-risk 5%, Moderate 17%, High 29%, Very High 46%; $p<0.0001$) (**Figure 2, panel B**). At Cox regression, the DL-CXR risk score was the strongest predictor of follow-up mortality ($p<0.005$; HR: 3.30; CI: 2.35-4.64) (**Figure 3**) and had an AUC of 0.793 (95% AUC CI: 0.759-0.827, Sens. 78%, Spec. 68%), that was significantly better than that achieved when using coronary angiography findings (AUC: 0.569, 95% CI: 0.526-0.611) ($p<0.001$) and age (AUC 0.735, 95% CI: 0.694-0.776) ($p<0.004$) as predictors. The DL-CXR

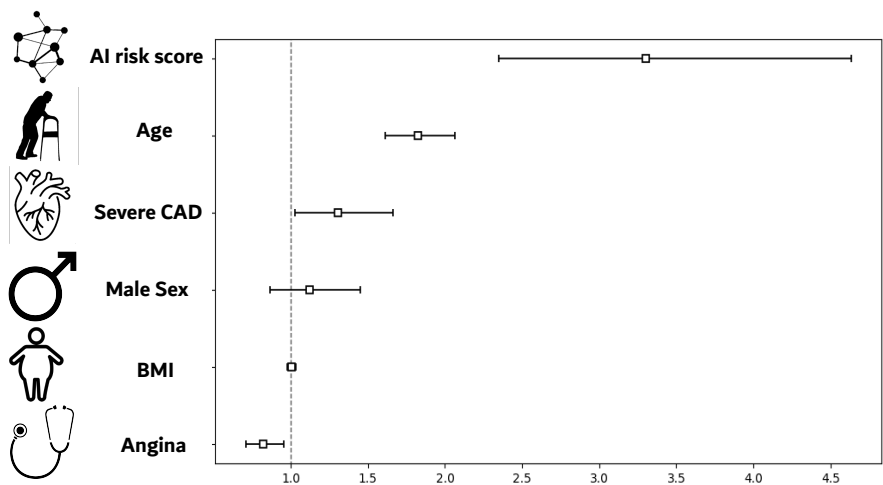


Figure 3. Forest plot illustrating hazard ratios (HR) for mortality.

mortality prediction model improved when age and CAD status at the referral time were included (AUC 0.81; CI: 0.775-0.843) (**Figure 4**).

Table 2 summarizes the confusion matrix of the model including CXR-score, age, and CAD status.

External validation with 193 patients confirmed the DL-CXR risk score prediction performance (AUC 0.66 CI: 0.495-0.928; Sens 0.667; Spec 0.606. KM survival analysis revealed a graded association between DL-CXR risk score categories and mortality ($p=0.04$) (**Supplementary Figure 1**).

Heat Maps: Attention maps were created considering the DCNN with the highest performance (i.e., the one including the DL-CXR risk score, CAD status and age). Heat map activations are primarily localized to the cardiac silhouette, left ventricular apex, pulmonary bases, pulmonary parenchyma, costophrenic sinuses, pulmonary hila, thoracic aorta, neck/supra-aortic vessels, and clavicle region (**Figure 5**).

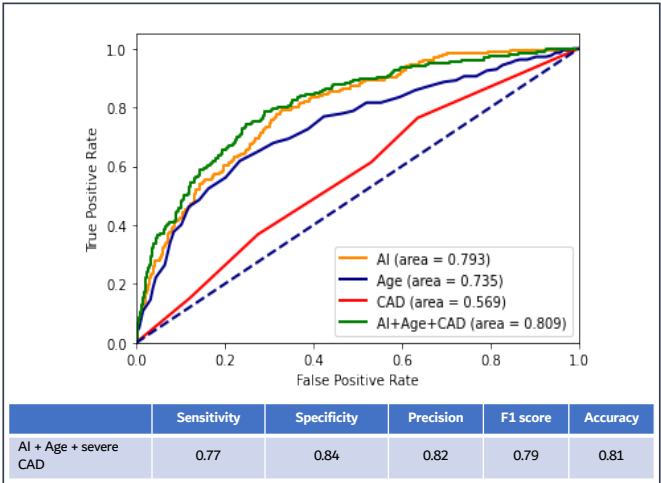


Figure 4. Receiver operating characteristic (ROC) curves for binary classification of survival/mortality. The area under the curve (AUC) values are as follows: 0.81 for AI + Age + CAD, 0.79 for DL-AI, 0.74 for Age, and 0.57 for CAD.

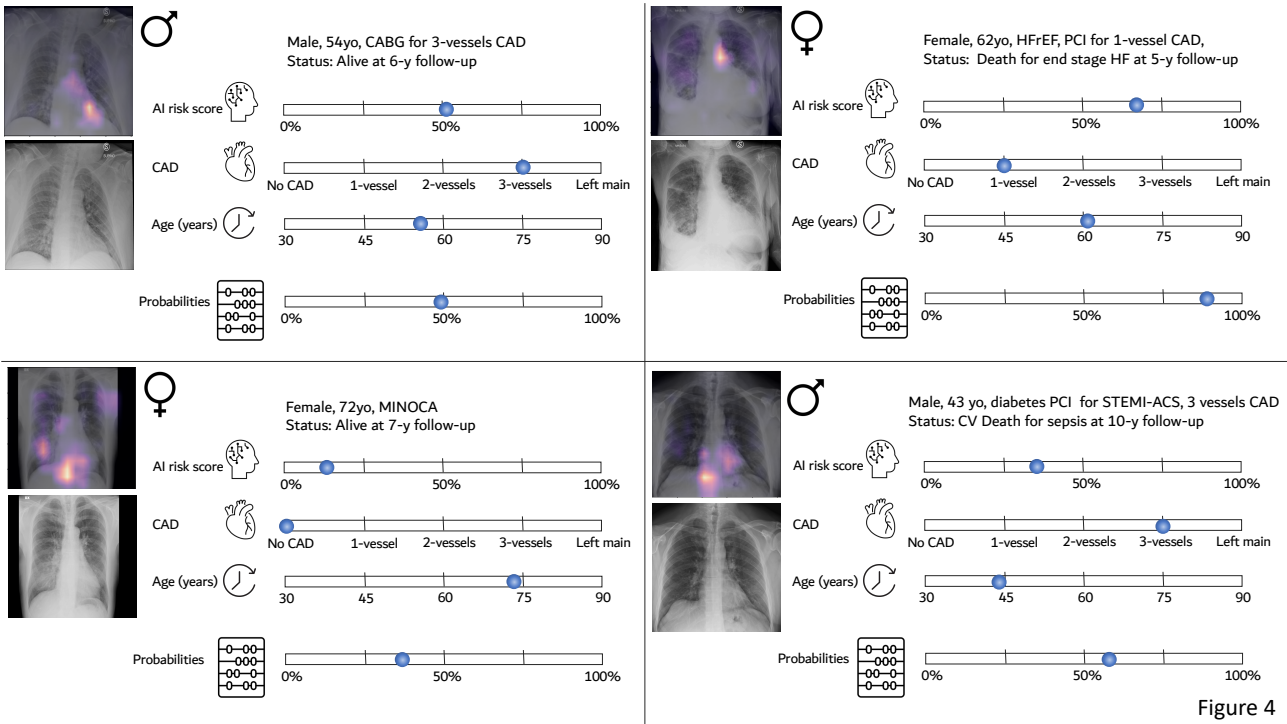


Figure 4

Figure 5. Heatmaps of four patients (two females and two males) showing areas of attention by AI in predicting mortality, highlighted in varying shades of yellow.

Discussion: Our study shows that: 1) the DCNN model predicted long-term (6-year) overall mortality from a single CXR view performed in patients with suspected angina pectoris referred for coronary angiography; 2) DL-based prediction of long-term mortality outperformed the prediction based on age, sex and CAD presence/severity; 3) The intrinsic generalizability of the model is supported by the fact that the dataset covered years of clinical routine, in a large hospital setting, where multiple radiographic acquisitions occurred using different radiographic equipment; 4) The performance of our DCNN has been confirmed at internal and external validation on patients referred for a similar reason (chest pain, CAD suspicion) to a different medical facility in a different European country; 5) Although our results are encouraging, the DCNN has not been tested on a broader population without chest pain, with a lower prevalence of CAD, with a less complex comorbidity profile who may have a lower long-term mortality rate.

CXR is a well-established and broadly used first-line diagnostic tool in patients with suspected cardiopulmonary conditions. It contains a plethora of information about the patient's general health, such as body habitus, cardiovascular condition, pulmonary status (lung congestion, interstitial lung disease, chronic obstructive pulmonary disease), and the bones/skeletal condition, including mineral density. All this information is related to the presence of comorbidities and a frailty profile that may impact follow-up survival [15-19] and may support the fact that an DL-read simple CXR may predict patients' outcomes better than any single risk factor taken alone.

We have previously demonstrated that DL-read CXRs could be used to pre-test significant CAD probability in patients referred for suspected angina [6]. A DCNN adequately trained in identifying severe CAD, using coronary angiography as ground truth, can detect the presence of severe CAD from a one-projection CXR and distinguish among patients with different degrees of CAD severity [6]. Starting from this assumption, we have focused the present study on the same patient cohort previously referred for coronary angiography and have trained, this time, our DCNN to predict mortality outcomes at long-term follow-up. In a prognostic study, Lu et al. have used a large cohort of patients (over 50,000) from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) to train and validate a DCNN to predict long-term mortality from chest radiograms. System performance was externally validated using data from the screening radiography arm of the National Lung Screening Trial (NLST) [5]. Using DCNN interpretation of diagnostic findings and standard risk factors, Lu et al. could identify patients with an increased risk of 6- and 12-year mortality. The CXR-risk score exhibited a graded association with mortality, ranging from 2.7% in the low-risk group to 33.9% in the very high-risk group [5]. The DL-CXR risk score achieved AUC values of 0.75 for 12-year mortality in the internal testing set and 0.68 for 6-year mortality in the external testing set [5]. In the same patient population, the same group has more recently confirmed that the DL CXR-

risk has a graded association with lung disease mortality after adjustment for risk factors, including age, smoking, and radiologic findings (with Hazard ratios up to 11) [20]. Our study, conducted in a cohort with a high prevalence of severe CAD (over 50%), further confirms this graded association. All quartiles of DL-CXR risk scores showed higher mortality rates, possibly due to the complex comorbid profile of our patients, particularly those with severe CAD, which led to poorer outcomes compared to patients in a cancer prevention program.

As emerging from our analysis and previously confirmed by Lu et al., the DCNN interpretation of the CXRs allows us to identify a prognostic value that is complementary to broadly used risk factors for follow-up mortality [6]. It has already been proposed that substituting biological age measures for chronological age could improve the performance of existing risk scores [21]. Raghu et al. and more recently Mitsuyama et al. have proposed an AI/DL-based estimate of biological age starting from a CXR [22, 23]. The CXR age predicts longevity and the presence of various chronic diseases beyond chronological age [22, 23]. There are possibly some common activation areas that can be identified within the radiographic image, like the mediastinum width, the aortic knob calcification and aortic tortuosity, the cardiac silhouette geometrical ratios, the bone and muscle density, and so on that modify with age. All these parameters could actually be used as surrogates for biological age and act together, to support a single estimator of the biological age and overall health status, for the AI-based prediction of chronic conditions, outcomes after interventions, and long-term events [24-29]. Our logistic regression and ROC statistics findings confirm that the AI/DL-based CXR-score can predict long-term mortality significantly better than chronological age.

Apart from the prediction of long-term outcomes, DL evaluation of emergency department CXRs has been used to identify patients referred for acute chest pain syndrome who are at risk for a 30-day composite adverse outcome, including acute coronary syndrome, pulmonary embolism, aortic dissection, and overall mortality [30]. The DL tool improved the prediction of these acute adverse outcomes beyond age, sex, and conventional troponin/d-dimer positivity [30]. In a first referral setting, the CXR-score could help clinicians to defer unnecessary additional, more invasive investigations in 14% of the referred patients [30].

It should be emphasized that the DL-based CXR score should not be extrapolated from the overall clinical setting, and additional clinical information concerning the patient's demography and health status may help optimize the image-based AI/DL prediction of events in the acute setting and during long-term follow-up. Kolossváry et al. have reported a significant improvement in the prediction of 30-day cardiovascular adverse events in a model that included, simultaneously, age, sex, biomarkers (troponin), and AI-based CXR score (from an AUC of 0.80 in the sole CXR-score model to an AUC of 0.85 in the CXR-score, sex, age, troponin model) [30]. Raghu et al. have shown only a slight

improvement in the prediction of 12-year follow-up mortality, from an AUC of 0.75 to an AUC of 0.78, when adding to the CXR-score based model information concerning documented risk factors (age, sex, smoking category, diabetes, hypertension, obesity, underweight, past myocardial infarction, past stroke, and past cancer) and reported CXR pathologic findings (such as lung nodule, significant atelectasis, pleural plaque or effusion, lymphadenopathy, chest wall or bony lesion, COPD/emphysema, cardiomegaly or other cardiovascular abnormality, and lung fibrosis) [22]. The impact of known pathologic conditions, such as CAD, on follow-up mortality has been documented and should be considered when designing DL-based prediction models [23]. All patients included in our cohort had undergone coronary angiography and could be stratified by the presence and extension of CAD. Interestingly, the CXR-based mortality prediction seemed more substantial than the CAD classification alone. The impact of CAD on mortality could have been indeed buffered, as all patients in the study underwent appropriate initiation and maintenance of adequate treatment after coronary angiography, including invasive interventions, whenever required [31] and, as a result, adding age and CAD status only slightly improved the performance of the CXR score.

Previously proposed mortality risk scores, built on linear models (such as logistic and linear regression), explainable tree-based models, and more complex DL models, have shown good accuracy, with AUCs reaching 0.90 in the prediction of long-term mortality, by including many demographics, laboratory, questionnaires, and at times epigenetic and gene expression data for the identification of so-called biological clocks [32-36].

Although we believe that DL models that incorporate additional demographic and clinical data, including multiple additional imaging and laboratory tests, may have a more accurate prognostic value and may be able to detect changes in risk that can occur at different time points and result from external human-driven interventions, we should be aware of the budgetary and organizational costs vs. accuracy trade-offs. In this light, the main aim of our research was to support the fact that the DCNN can extract from a simple, affordable, and ubiquitous investigation, such as the CXR performed at the time of referral in different settings, the "hidden fingerprints" of long-term outcomes that are embedded in the image, and that the DCNN performance is independent of additional demographic or clinical information.

Limitations

The DCNN has been trained, tested, and validated on patients referred for invasive coronary angiography, considering a clinical suspect of angina. Critical patients' selection biases limit our findings. The population carries a specific burden of comorbidities that is not necessarily present in other patients' groups and may have impacted events occurrence. For this reason, the DCNN should

also be validated in a broader population of healthy people that are asymptomatic for chest pain. Moreover, because of the study's retrospective nature and many additional budgetary/logistical limitations, we have included only a limited amount of clinical information from our study cohort. In the future, to achieve a more accurate and personalized non-static event prediction and clinical utility of the model, the CXR-score will need to be integrated with additional demographic, clinical, and possibly medical imaging, laboratory test findings, time-dependent events, and human interventions. Those implementations may impact the model's explainability, sustainability, and fairness (applicability in different and less structured healthcare environments). Finally, shedding light on the "black box" of the DL algorithm's decision-making process is not an easy task, and the suboptimal explainability of the model may hinder its application in the healthcare sector. Although the produced attention maps may partly enhance the CXR areas most often involved in the mortality prediction, we still do not know why the model assigns a specific score and makes a particular prediction. We will need to address this issue in the future by investigating the cause of death and resorting to techniques around explainable AI/DL that help clarify complex DL models.

Conclusion

CXRs are routinely performed in patients referred for different conditions and healthy subjects for screening. This simple and affordable imaging modality should be considered for risk stratification, including the prediction of long-term mortality. Our DCNN model predicted, from a CXR, long-term overall mortality of patients with suspected angina pectoris referred for coronary angiography. The performance has been confirmed at internal and external validation on patients in two different European regions (north and south). The DCNN CXR-risk score could stratify long-term mortality risk better than any other risk factor, including CAD presence and extension, sex, and chronological age. These findings may be used to optimize screening and prevention actions for this very comorbid group of patients, although at present, our findings cannot be extended to patients' cohorts with different risk profiles, including healthy individuals. In the future, the DL-based CXR-score could be used as a supportive, easy-to-apply tool to estimate, together with a plethora of additional tests, the risk of follow-up adverse events and to guide, in this way, anticipated actions to change unhealthy lifestyles and trigger a prompt referral to additional and more specific investigations, whenever clinically sound.

References

- 1) Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, Prescott E, Storey RF, Deaton C, Cuisset T, Agewall S, Dickstein K, Edvardsen T, Escaned J, Gersh BJ, Svitil P, Gilard M, Hasdai D, Hatala R, Mahfoud F, Masip J, Muneretto C, Valgimigli M, Achenbach S, Bax JJ. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. *Eur Heart J*; 2020;**41**:407–477.
- 2) Ron E. Cancer risks from medical radiation. *Health Phys*. 2003;**85**:47-59.
- 3) Bader AS, Rubinowitz AN, Gange CPJ, Bader EM, Cortopassi IO. Imaging in the Evaluation of Chest Pain in the Primary Care Setting, Part 1: Cardiovascular Etiologies. *Am J Med*; 2020;**133**:1033–1038.
- 4) Seah JCY, Tang CHM, Buchlak QD, Holt XG, Wardman JB, Aimoldin A, Esmaili N, Ahmad H, Pham H, Lambert JF, Hachey B, Hogg SJF, Johnston BP, Bennett C, Oakden-Rayner L, Brothie P, Jones CM. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Heal*; 2021;**3**:e496–e506.
- 5) Lu MT, Ivanov A, Mayrhofer T, Hosny A, Aerts HJWL, Hoffmann U. Deep Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA Netw open* 2019;**2**:e197416.
- 6) D'Ancona G, Massucci M, Savardi M, Signoroni A, Di Bacco L, Farina D, Metra M, Maroldi R, Muneretto C, Ince H, Costabile D, Murero M, Chizzola G, Curello S, Benussi S. Deep learning to detect significant coronary artery disease from plain chest radiographs AI4CAD. *Int J Cardiol*. 2022 Nov 5:S0167-5273(22)01665-5.
- 7) Sexs TSS, Steyerberg EW, Alkadhi H, Leschka S, Desbiolles L, Nieman K, Galema TW, Meijboom WB, Mollet NR, Feyter PJ de, Cademartiri F, Maffei E, Dewey M, Zimmermann E, Laule M, Pugliese F, Barbagallo R, Sinitsyn V, Bogaert J, Goetschalckx K, Schoepf UJ, Rowe GW, Schuijf JD, Bax JJ, Graaf FR de, Knuuti J, Kajander S, Mieghem CAG van, Meijs MFL, Cramer MJ, et al. A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. *Eur Heart J*; 2011;**32**:1316–1330.
- 8) Bustos A, Pertusa A, Salinas J-M, de la Iglesia-Vayá M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 2020; **66**:101797.
- 9) Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017; 3462-3471.
- 10) Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghighi B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *AAAI Conference on Artificial Intelligence* 2019; **33**(01):590-597.
- 11) Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-y, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019; **6**:317.
- 12) He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)* 2016; 770-778.
- 13) Cohen JP, Viviano JD, Bertin P, Morrison P, Torabian P, Guarrera M, Lungren MP, Chaudhari A, Brooks R, Hashir M, Bertrand H. TorchXRyVision: A library of chest X-ray datasets and models. *arXiv preprint* 2021; arXiv:2111.00595.
- 14) Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)* 2017. p. 618–626.
- 15) Clarson LE, Bajpai R, Whittle R, Belcher J, Abdul Sultan A, Kwok CS, Welsh V, Mamas M,

- Mallen CD. Interstitial lung disease is a risk factor for ischaemic heart disease and myocardial infarction. *Heart* 2020;**106**:916–922.
- 16) Boschetto P, Beghé B, Fabbri LM, Ceconi C. Link between chronic obstructive pulmonary disease and coronary artery disease: implication for clinical practice. *Respirology Australia*; 2012;**17**:422–431.
 - 17) Zhang Y, He B, Wang H, Shi J, Liang H. Associations between bone mineral density and coronary artery disease: a meta-analysis of cross-sectional studies. *Arch Osteoporos England*; 2020;**15**:24.
 - 18) Kumar DA, Anburajan M. The role of hip and chest radiographs in osteoporotic evaluation among south Indian women population: a comparative scenario with DXA. *J Endocrinol Invest Italy*; 2014;**37**:429–440.
 - 19) Zaman MJS, Sanders J, Crook AM, Feder G, Shipley M, Timmis A, Hemingway H. Cardiothoracic ratio within the ‘normal’ range independently predicts mortality in patients undergoing coronary angiography. *Heart* 2007;**93**:491–494.
 - 20) Weiss J, Raghu VK, Bontempi D, Christiani DC, Mak RH, Lu MT, Aerts HJWL. Deep learning to estimate lung disease mortality from chest radiographs. *Nat Commun*. 2023;14(1):2797.
 - 21) Grundy SM. Coronary plaque as a replacement for age as a risk factor in global risk assessment. *Am J Cardiol* 2001;88:8E–11E.
 - 22) Raghu VK, Weiss J, Hoffmann U, Aerts HJWL, Lu MT. Deep Learning to Estimate Biological Age From Chest Radiographs. *JACC Cardiovasc Imaging*. 2021;14:2226–2236. doi: 10.1016/j.jcmg.2021.01.008. Epub 2021 Mar 17.
 - 23) Mitsuyama Y, Matsumoto T, Tatekawa H, Walston SL, Kimura T, Yamamoto A, Watanabe T, Miki Y, Ueda D. Chest radiography as a biomarker of ageing: artificial intelligence-based, multi-institutional model development and validation in Japan. *Lancet Healthy Longev* 2023; S2666-7568(23)00133-2.
 - 24) Pyrros A, Borstelmann SM, Mantravadi R, Zaiman Z, Thomas K, Price B, Greenstein E, Siddiqui N, Willis M, Shulhan I, Hines-Shah J, Horowitz JM, Nikolaidis P, Lungren MP, Rodríguez-Fernández JM, Gichoya JW, Koyejo S, Flanders AE, Khandwala N, Gupta A, Garrett JW, Cohen JP, Layden BT, Pickhardt PJ, Galanter W. Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. *Nat Commun*. 2023;14:4039.
 - 25) Ueda D, Matsumoto T, Ehara S, Yamamoto A, Walston SL, Ito A, Shimono T, Shiba M, Takeshita T, Fukuda D, Miki Y. Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. *Lancet Digit Health*. 2023;5:e525–e533.
 - 26) Ryu J, Eom S, Kim HC, Kim CO, Rhee Y, You SC, Hong N. Chest X-ray-based opportunistic screening of sarcopenia using deep learning. *J Cachexia Sarcopenia Muscle*. 2023;14:418–428.
 - 27) Raghu VK, Moonsamy P, Sundt TM, Ong CS, Singh S, Cheng A, Hou M, Denning L, Gleason TG, Aguirre AD, Lu MT. Deep Learning to Predict Mortality After Cardiothoracic Surgery Using Preoperative Chest Radiographs. *Ann Thorac Surg*. 2023;115:257–264.
 - 28) Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N. Opportunistic Osteoporosis Screening Using Chest Radiographs With Deep Learning: Development and External Validation With a Cohort Dataset. *J Bone Miner Res*. 2022;37:369–377.
 - 29) Ieki H, Ito K, Saji M, Kawakami R, Nagatomo Y, Takada K, Kariyasu T, Machida H, Koyama S, Yoshida H, Kurosawa R, Matsunaga H, Miyazawa K, Ozaki K, Onouchi Y, Katsushika S, Matsuoka R, Shinohara H, Yamaguchi T, Kodera S, Higashikuni Y, Fujiu K, Akazawa H, Iguchi N, Isobe M, Yoshikawa T, Komuro I. Deep learning-based age estimation from chest X-rays indicates cardiovascular prognosis. *Commun Med*. 2022;2:159.
 - 30) Kolossváry M, Raghu VK, Nagurney JT, Hoffmann U, Lu MT. Deep Learning Analysis of Chest Radiographs to Triage Patients with Acute Chest Pain Syndrome.

Radiology. 2023;306:e221926. doi: 10.1148/radiol.221926.

- 31) Hanson CA, Lu E, Ghumman SS, Ouellette ML, Löffler AI, Beller GA, Bourque JM. Long-term outcomes in patients with normal coronary arteries, nonobstructive, or obstructive coronary artery disease on invasive coronary angiography. Clin Cardiol. 2021 Sep;44(9):1286-1295. doi: 10.1002/clc.23686.
- 32) Qiu W, Chen H, Dincer AB, Lundberg S, Kaeberlein M, Lee SI. Interpretable machine learning prediction of all-cause mortality. Commun Med 2022; 2:125. doi: 10.1038/s43856-022-00180-x
- 33) Ganna A, Ingelsson E. 5 year mortality predictors in 498 103 UK Bioparticipants: A prospective population-based study. Lancet 2015; 386, 533–540
- 34) Horne BD, May HT, Muhlestein JB, Ronnow BS, Lappé DL, Renlund DG, Kfoury AG, Carlquist JF, Fisher PW, Pearson RR, Bair TL, Anderson JL. Exceptional mortality prediction by risk scores from common laboratory tests. Am J Med 2009;122:550-8. doi: 10.1016/j.amjmed.2008.10.043.
- 35) Liu Z, Kuo PL, Horvath S, Crimmins E, Ferrucci L, Levine M. A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: A cohort study. PLoS Med. 2018;15:e1002718. doi:10.1371/journal.pmed.1002718.
- 36) Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018;19:371-384. doi: 10.1038/s41576-018-0004-3.

Table 1. Baseline Characteristics of the Overall Study Population and of the Internal Validation Cohort Split into 4 Quartiles According to DL-CXR Risk Score.

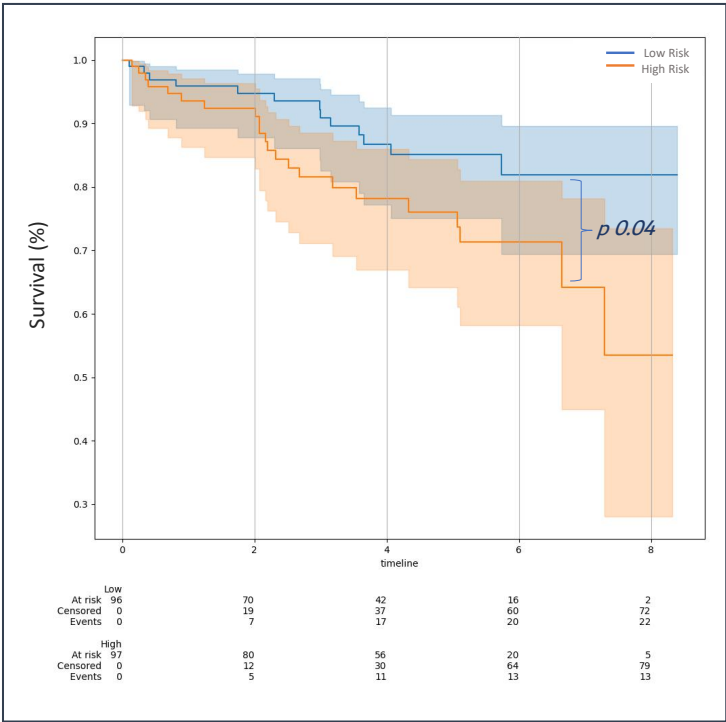
	Overall Population (N=6031)	Validation cohort (N=1170)			
		Low Risk (N=234)	Moderate Risk (N=206)	High Risk (N=318)	Very High Risk (N=412)
Age, y	68±12	69±12	71±11	67±11	64±12
Biological Sex (M)	67%	65%	63%	69%	75%
BMI	26±5	26±5	26±5	26±4	27±4
Coronary angiography					
Non-obstructive CAD	44%	41%	46%	44%	47%
1 Vessel CAD	27%	26%	29%	28%	23%
2 Vessels CAD	16%	18%	15%	18%	17%
3 Vessels CAD	12%	14%	11%	10%	13%
Left Main CAD	1%	1%	1%	0%	0%

BMI=body mass index; CAD=coronary artery disease

Table 2. Confusion matrix and extracted metrics of the internal validation cohort

Predicted label				
True label	129		123	
	84		834	
AUC	Sensitivity	Specificity	Precision	NPV
0.81	87%	61%	91%	49%

Supplementary Figure. External validation of DL-CXR risk score in the overall population, with the two groups stratified by AI prediction (A), and corresponding ROC curves with an AUC of 0.66 (B)





Deep learning to predict long-term mortality from plain chest X-ray in patients referred for suspected coronary artery disease

Giuseppe D'Ancona¹, Mattia Savardi^{2,3}, Mauro Massussi⁴, Viktor Van Der Valk⁵, Roderick W. C. Scherptong⁵, Alberto Signoroni^{2,3}, Davide Farina⁶, Monica Murero⁷, Hüseyin Ince¹, Stefano Benussi⁸, Salvatore Currello⁴, Fatih Arslan⁹

¹Department of Cardiology and Cardiovascular Clinical Research Unit, Vivantes Klinikum Urban and Neukölln, Berlin, Germany; ²Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, Brescia, Italy; ³Department of Information Engineering, University of Brescia, Brescia, Italy; ⁴Cardiac Catheterization Laboratory, Department of Cardiothoracic, ASST Spedali Civili, Brescia, Italy; ⁵Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands; ⁶Radiology 2, ASST Spedali Civili and Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, Brescia, Italy; ⁷Department of Excellence in Social Sciences, University Federico II, Naples, Italy; ⁸Department of Cardiac Surgery, Spedali Civili Brescia and University of Brescia, Brescia, Italy; ⁹Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands

Contributions: (I) Conception and design: G D'Ancona, M Massussi, S Currello; (II) Administrative support: A Signoroni, S Currello, D Farina; (III) Provision of study materials or patients: M Massussi, S Currello, D Farina, F Arslan, RWC Scherptong; (IV) Collection and assembly of data: M Massussi, M Savardi, F Arslan, V Van Der Valk; (V) Data analysis and interpretation: M Savardi, A Signoroni, G D'Ancona, M Massussi, V Van Der Valk, F Arslan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Giuseppe D'Ancona, MD, PhD. Department of Cardiology and Cardiovascular Clinical Research Unit, Vivantes Klinikum Am Urban, Diefenbachstraße 1, 10967 Berlin, Germany. Email: rgea@hotmail.com.

Background: The hypothesis that a deep learning (DL) model can produce long-term prognostic information from chest X-ray (CXR) has already been confirmed within cancer screening programs. We summarize our experience with DL prediction of long-term mortality, from plain CXR, in patients referred for angina and coronary angiography.

Methods: Data of patients referred to an Italian academic hospital were analyzed retrospectively. We designed a deep convolutional neural network (DCNN) that, from CXR, could predict long-term mortality. External validation was performed on patients referred to a Dutch academic hospital.

Results: A total of 6,031 were used for model training (71%; n=4,259) and fine-tuning/validation (10%; n=602). Internal validation was performed with the remaining patients (19%; n=1,170). Patients' stratification followed the DL-CXR risk score quartiles division. Median follow-up was 6.1 years [interquartile range (IQR), 3.3–8.7 years]. We observed an increment in estimated mortality with the increase of DL-CXR risk score (low-risk 5%, moderate 17%, high 29%, very high 46%; $P<0.001$). The DL-CXR risk score predicted median follow-up outcome with an area under the curve (AUC) of 0.793 [95% confidence interval (CI): 0.759–0.827, sensitivity 78%, specificity 68%]. Prediction was better than that achieved using coronary angiography findings (AUC: 0.569, 95% CI: 0.52–0.61, $P<0.001$) and age (AUC: 0.735, 95% CI: 0.69–0.77, $P<0.004$). At Cox regression, the DL-CXR risk score predicted follow-up mortality ($P<0.005$, hazard ratio: 3.30, 95% CI: 2.35–4.64). External validation confirmed the DL-CXR risk score performance (AUC: 0.71, 95% CI: 0.49–0.92; sensitivity 0.838; specificity 0.338).

Conclusions: In patients referred for coronary angiogram because of angina, the DL-CXR risk score could be used to stratify mortality risk and predict long-term outcome better than age and coronary artery disease status.

Keywords: Machine learning; chest X-ray (CXR); mortality; angina

3. Use Case 2: Self-Supervised Learning for ECG Interpretation (HUBERT-ECG Foundation Model)

3.1 Rationale and Innovation: The Foundation Model Paradigm for ECG Interpretation

The electrocardiogram (ECG) is one of the most widely used diagnostic tools in cardiovascular medicine, providing a rapid, inexpensive, and non-invasive assessment of cardiac physiology. For over a century, clinicians have relied on the ECG to detect arrhythmias, conduction abnormalities, ischemia, and structural heart disease. Despite its fundamental role, ECG interpretation demands considerable expertise, and even experienced clinicians are limited by the physiological complexity of electrical signals and by inter-reader variability¹⁹. In recent years, AI has begun to reshape the way clinicians interpret this ubiquitous test, revealing that the ECG contains far more diagnostic and prognostic information than previously recognized.

AI—particularly deep learning—has opened the possibility of decoding subtle patterns hidden within raw ECG waveforms, many of which are beyond the perceptual threshold of human readers. Convolutional and transformer-based architectures have already demonstrated the capability to reproduce human-like ECG interpretation at scale²⁰, but also to identify conditions not classically diagnosable from surface ECGs. Large linked datasets of ECGs and clinical outcomes have enabled models to detect left ventricular systolic dysfunction, silent atrial fibrillation, hypertrophic cardiomyopathy, electrolyte disturbances, and even demographic traits such as age and sex²¹. These findings have established the ECG as a powerful, fully digital biomarker that can support a wide variety of clinical predictions.

The ECG is an ideal substrate for advanced AI for several reasons. It is performed millions of times per day worldwide, requires no specialized equipment beyond standard acquisition devices, and produces uniform digital signals that are highly reproducible and easily transferable. The growing availability of cloud-based ECG repositories and wearable ECG devices further expands the potential of AI-enhanced ECGs (AI-ECGs), enabling their integration into routine care, telemedicine pathways, and population-level screening programs. As research in AI-ECG evolves, it has become increasingly clear that these models can extract a multidimensional fingerprint of cardiovascular health, capturing information on cardiac structure, conduction properties, autonomic tone, and even systemic conditions.

However, despite these advances, current AI-ECG systems are usually trained for narrow, task-specific problems. Most models are designed to detect a single disease or predict one outcome at a

time²². This fragmentation limits their generalizability across patient populations, healthcare settings, and clinical tasks. Moreover, the creation of high-performing models typically requires large datasets that are not always available, especially for rare diseases or smaller clinical centers. These limitations have motivated the search for a new paradigm: the development of foundation models for ECG analysis.

Foundation models—large-scale architectures trained on massive unlabeled datasets using self-supervised learning—have transformed natural language processing and computer vision, providing models that can be efficiently adapted to many downstream tasks with minimal fine-tuning²³. Translating this paradigm to cardiovascular medicine offers the potential to unify ECG interpretation into a single, versatile model capable of supporting a wide range of diagnostic and prognostic applications.

The clinical logic behind an ECG-based foundation model is clear. The ECG can inform three major categories of cardiovascular assessment:

1. Diseases for which the ECG is the primary diagnostic tool. These include arrhythmias, conduction disturbances, channelopathies, and several electrical or repolarization disorders. A foundation model trained on millions of ECGs could replicate and enhance clinician-level expertise for these conditions, enabling rapid triage, screening, and quality control across diverse healthcare settings.

2. Diseases where the ECG provides supportive—but not definitive—diagnostic information. In conditions such as heart failure, valvular disease, pulmonary hypertension, ischemia, and cardiomyopathies, the ECG reflects secondary electrical or structural changes that are clinically informative but often subtle. AI can amplify these weak signals, allowing for improved detection or risk stratification even when the ECG is not the primary diagnostic modality.

3. Prediction of cardiovascular events and outcomes. Perhaps the most disruptive potential lies in prognostication. Deep learning has shown that ECG waveform patterns encode information about long-term mortality, cardiometabolic risk, vascular aging, and incident disease²⁴. A foundation model could unify these capabilities, enabling precise forecasting of adverse events directly from a routine ECG.

This clinical framework—spanning diagnostic, supportive, and prognostic applications—motivates the development of a generalizable, high-capacity model that can be adapted to a multitude of real-world tasks. In this context, we introduce HuBERT-ECG, a self-supervised, large-scale foundation model pre-trained on an unprecedented dataset of 9.1 million 12-lead ECGs, encompassing 164 cardiovascular conditions across diverse populations and care settings. By adding simple task-specific output layers, HuBERT-ECG can be fine-tuned for a broad array of applications ranging from disease

detection to event prediction. In 17 real-world downstream tasks, the model achieves AUROCs ranging from 0.843 in small datasets to 0.99 in large clinical cohorts, demonstrating exceptional robustness. When trained to detect 164 overlapping conditions, it achieves AUROCs above 0.90 for 140 diseases and above 0.95 for 97 diseases. Moreover, HuBERT-ECG predicts 2-year mortality with AUROCs up to 0.91, confirming its ability to capture prognostic signatures embedded in the ECG. HuBERT-ECG is therefore designed not as a single-purpose model, but as a clinically grounded, generalizable, and scalable ECG foundation model capable of supporting three major roles: high-accuracy diagnosis for diseases fundamentally encoded in the ECG, enhanced detection of structural and systemic conditions where ECG findings are subtle, and powerful prediction of adverse cardiovascular outcomes. By releasing both the pre-trained model and the accompanying codebase, this work establishes a strong baseline for future research and clinical translation. Ultimately, HuBERT-ECG represents a meaningful step toward unifying AI-ECG applications under a single, flexible model, paving the way for a new generation of AI-enabled cardiovascular diagnostics and risk prediction.

3.2 HuBERT-ECG: Self-Supervised Learning for Scalable, Multitask ECG Diagnosis and Prediction

HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications

Edoardo Coppola, Mattia Savardi, Mauro Massucci, Marianna Adamo, Marco Metra, Alberto Signoroni

ABSTRACT

Deep learning models have shown remarkable performance in electrocardiogram (ECG) analysis, but the limited availability and size of ECG datasets have constrained their success, resulting in systems that are more task specialists than versatile generalists. To counter this, we introduce HuBERT-ECG, a novel self-supervised foundation ECG model pre-trained on a large and diverse dataset of 9.1 million 12-lead ECGs encompassing 164 cardiovascular conditions. By simply adding a proper output layer, HuBERT-ECG can be fine-tuned for a wide array of downstream tasks, from diagnosing diseases to predicting future cardiovascular events. Across 17 diverse real-world scenarios, HuBERT-ECG achieves AUROCs from 0.843 on small datasets to 0.99 on larger sources. When fine-tuned to detect 164 overlapping conditions simultaneously, our model delivers AUROCs above 0.9 and 0.95 for up to 140 and 97 diseases, respectively. HuBERT-ECG can also predict death events within a 2-year follow-up with AUROCs up to 0.91. We release pre-trained models and code as building baselines.

Introduction

The electrocardiogram (ECG) has long been a cornerstone of cardiovascular diagnostics, serving as a non-invasive, cost-effective and widely available tool for assessing cardiac and noncardiac diseases¹. Through its recurring waveforms, the ECG captures a unique “language of the heart”, encoding both physiological and pathological vital information through distinctive electrophysiological “fingerprints”². This language, although seemingly simple in its structure, reveals complex and meaningful patterns that allow the identification of deviations from expected cardiac functions, providing valuable insights into heart health.

Integrating deep learning (DL) into ECG analysis represents a transformational opportunity to significantly improve its clinical utility³. In fact, DL algorithms can transform raw ECG traces into powerful digital biomarkers, enabling early detection, risk stratification, and intervention across a wide range of cardiovascular conditions. Previous studies have demonstrated substantial impact of DL in specific areas of ECG interpretation. For example, DL models have consistently shown solid

performance in detecting conditions where the ECG is the gold standard for diagnosis (e.g. atrial fibrillation, tachycardia, and bradycardia), leading to validated algorithms⁴. Conversely, DL applications in the diagnosis of morphological conditions, such as heart failure, pulmonary thromboembolism and aortic stenosis, are less widespread but growing⁵. Although these conditions rely primarily on imaging modalities, previous works demonstrate that DL can successfully use ECG data to predict cardiovascular conditions beyond those traditionally associated with ECG patterns, including non-cardiac conditions such as liver disease or diabetes⁶, providing a valuable complement to conventional imaging approaches⁷. Finally, the use of DL to predict future cardiovascular events (CVEs), such as myocardial infarction and stroke, represents the most ambitious application of this technology. While early results are promising, this area presents greater challenges due to the complexity of predicting outcomes from ECG patterns⁸. From a machine learning perspective, current limitations in ECG dataset size, availability, range of conditions and patient demographics have led to developing ECG models that are more task specialists⁹ rather than competent and versatile generalists, traits that typically characterise foundation models¹⁰. In fact, specialised models often struggle to adapt to new domains or even different data distributions within the same task¹¹. For example, an ECG model trained solely to detect atrial fibrillation on a dataset where each recording is labelled as either positive or negative for this condition will be unable to identify other cardiac abnormalities unless retrained on a new dataset. Similarly, if a model is trained exclusively on ECGs from individuals over 65, where atrial fibrillation is prevalent, it might fail to perform well on younger patients. In stark contrast, by pre-training in a self-supervised manner on large and diverse unlabelled datasets, foundation models can learn robust and generalised data representations that are transferable to a wide variety of downstream tasks, requiring only lightweight fine-tuning for specific domains or distributions¹².

Unimodal foundation models have achieved significant success across traditional domains—including natural language processing (e.g., the GPT series¹³⁻¹⁶, BERT¹⁷, T5¹⁸), computer vision (e.g., DINO¹⁹, MAE²⁰), and audio processing (e.g., Wav2Vec2²¹, HuBERT²²)—by leveraging large-scale self-supervised pre-training on extensive unannotated datasets. More recently, these advancements have also facilitated multimodal learning, where models integrate multiple data modalities to achieve a more holistic understanding of information within data (e.g., Google Gemini¹, GPT-4o²). In the medical domain, progress has been slower due to the limited availability of large medical datasets^{11,22}. However, this trend has begun to shift, leading to the emergence of foundation medical models such as CONCH²⁴, Prov-GigaPath²⁵ and Virchow²⁶ for computational pathology; KAD²⁷ and CheXzero²⁸

¹ <https://gemini.google.com/>

² <https://openai.com/index/hello-gpt-4o/>

for radiology; MedSAM²⁹, for medical image segmentation; EchoCLIP³⁰ for echocardiography; and BiomedGPT³¹ for various biomedical tasks. While recent efforts in electrocardiography attempted to develop foundation ECG models, they invariably lack one or more critical characteristics: large-scale and diverse pre-training data; self-supervised learning to learn generalisable representations; comprehensive evaluation across multiple datasets and tasks (particularly with scarce data); efficient fine-tuning for broad applicability. For instance, several works³²⁻³⁷ leverage contrastive-learning methods, but miss the combination of large-scale, diverse pre-training and extensive evaluation. HeartBEiT³⁸ exploits self-supervision during its large-scale pre-training, but suffers from limited source diversity and narrowly scoped evaluation, testing performance on only three tasks. KED³⁹ demonstrates impressive zero-shot capabilities via text supervision, but lacks large-scale, diverse pre-training data, omits self-supervision (reaching less general representations¹⁹) and is evaluated only on form and rhythm conditions, thereby limiting broad diagnostic utility. Most recently, ECGFounder⁴⁰ offers the ability to work for the identification of 150 conditions through a supervised training of a model on a large single-site database (Harvard Emory) of 10 million ECGs.

In this work, inspired by HuBERT²² architecture, we present HuBERT-ECG, a new ECG foundation model specifically designed for broad transferability and generalisation across a wide range of datasets and tasks, from diagnosing diseases to predicting future CVEs. HuBERT-ECG is pre-trained through self-supervision on an extensive dataset of 9.1 million ECGs from a large and diverse population across four countries. Noteworthy, over 2.4 million ECGs in our dataset are annotated with one or more conditions from a comprehensive set of 164 diagnoses, allowing HuBERT-ECG to learn and generalise over a wide spectrum of pathologies. These diagnostic labels are derived exclusively from the ECG signal—without integration of clinical, biomarker, or imaging data—and accurately reflect how ECGs are interpreted in current practice by automated systems or ECG-reading physicians. When fine-tuned and evaluated on the unparalleled number of 17 datasets, simulating numerous real-world scenarios of varying complexity, size and diversity, HuBERT-ECG demonstrates both efficiency and accuracy, showcasing the potential to serve as a strong foundation for addressing three crucial types of tasks: (1) detecting conditions where the ECG is the primary diagnostic tool (e.g., atrial fibrillation, bundle branch blocks); (2) supporting the identification of conditions where the ECG plays a non-primary but potentially informative role (e.g., heart failure, myocardial infarction); and (3) predicting future cardiovascular events (e.g., death, hospitalization). Finally, to support ongoing research on cardiovascular diseases and accelerate innovation, we release HuBERT-ECG models and code to the community. We make available HuBERT-ECG in three model sizes (or configurations)—SMALL (30 million parameters), BASE (93 million parameters) and LARGE (188 million parameters)—to both meet the computational constraints of future users and

address a broad range of applications where some model sizes may perform better than others. By sharing insights from comprehensive, benchmarked evaluations on many diverse downstream datasets, we aim to facilitate HuBERT-ECG successful adoption and customisation for research purposes.

Results

ECG Representation Learning through Self-supervised Pre-training

To develop a robust and generalisable ECG foundation model, we propose HuBERT-ECG, a self-supervised foundation model inspired by HuBERT²², originally designed for speech signals. HuBERT-ECG learns meaningful ECG representations through an iterative process combining (1) feature extraction, (2) clustering and (3) masked modelling. Specifically, descriptive ECG features are extracted from ECG segments and then clustered by a clustering model. Subsequently, during pre-training, embedded versions of ECG segments are masked and HuBERT-ECG is trained to predict which cluster they belong to. By repeating this process across multiple pre-training iterations, and

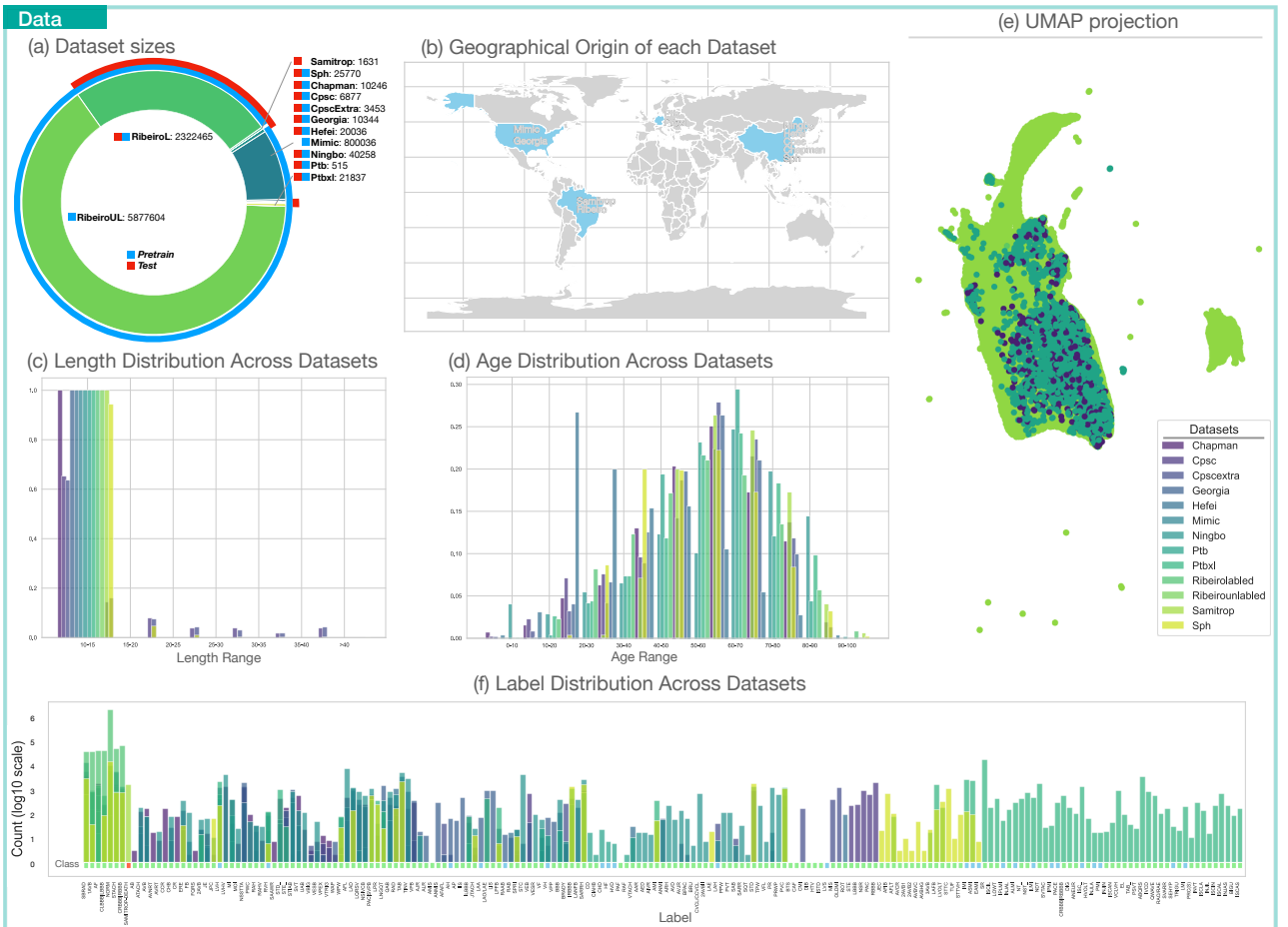


Figure 1. General information on the 9.1 million ECG dataset (a) Used dataset and their sizes (b) Geographical origin of the datasets. (c) ECG length (d) Age distribution (e) UMAP projection (uniform sample of 30% of all the dataset for the sake of visualisation). (f) Label distribution and grouping according to 3 classes

progressively refining the feature extraction, HuBERT-ECG improves its ability to capture the underlying structure of ECG signals, ultimately achieving generalisable ECG representations.

We pre-train HuBERT-ECG on an extensive and diverse dataset of 9.1 million ECGs from four countries (**Fig. 1a-d, Fig. 2a**, “Methods” – *Data and Preprocessing*).

In the first iteration, prior to masked modelling, we extract and cluster Mel Frequency Cepstral Coefficients⁴¹ (MFCCs) from raw ECG segments. MFCCs, known for capturing key frequency components of the signal, highlight the essential structure of the ECG. In the second iteration, we extract semantically higher features from the model’s 8th layer (latent representations) and refine the clusters. This iterative refinement allows HuBERT-ECG to learn a deeper understanding of ECG patterns and structure through masked modelling. HuBERT-ECG architecture and pre-training specifics are detailed in “Methods” - *HuBERT-ECG Architecture and Theoretical Framework*. Feature extraction and clustering are illustrated in “Methods” – *Unsupervised label discovery*. Specifications and implementation details are provided in “Methods” – *Implementations* and “Methods” – *Self-supervised pre-training*.

To further evaluate the adaptability of our approach and offer models capable of addressing diverse needs after fine-tuning, we also pre-train SMALL and LARGE model sizes in a single iteration using more refined latent representations.

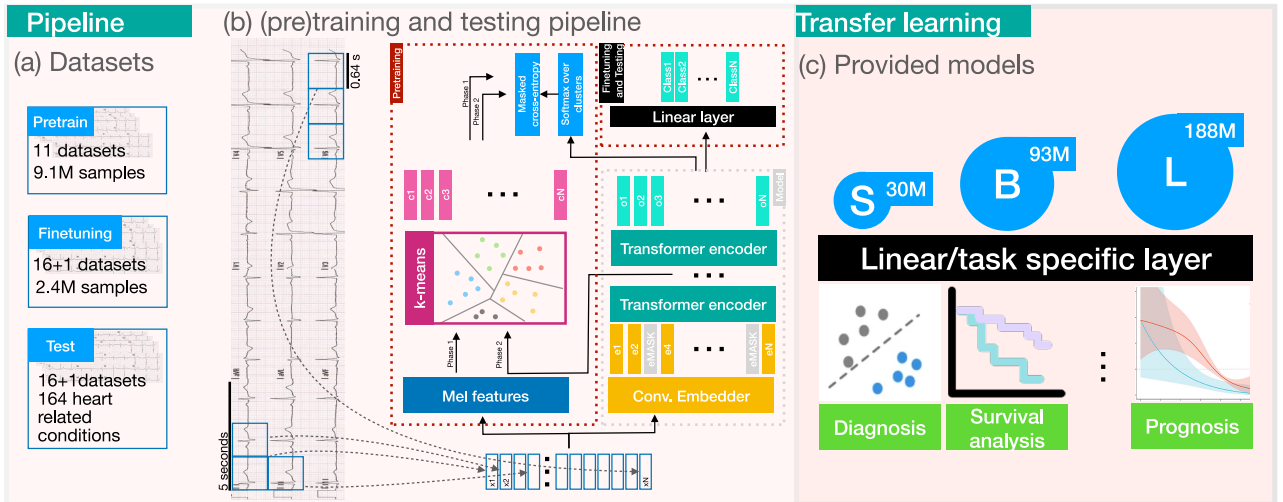


Figure 11. (a) Datasets: The model is pre-trained on 11 datasets comprising over 9.1 million samples. Then, the model is fine-tuned and tested on 16 datasets and their aggregation covering 164 heart-related conditions. (b) Pre-training and Testing Pipeline. (c) T Transfer Learning and Provided Models: Three model variants (**SMALL**, **BASE**, **LARGE**)

Supervised fine-tuning on downstream data and evaluation

To comprehensively assess HuBERT-ECG’s performance and adaptability across various clinical scenarios, we compile 16 downstream datasets for fine-tuning and evaluation. These datasets include: the labelled partition of Ribeiro⁴² (also known as CODE); CPSC and CPSC-Extra⁴³; PTB⁴⁴; PTB-XL⁴⁵ which, with 6 different sets of conditions, gives rise to 6 different datasets; the publicly available partition of Georgia⁴⁶; Chapman-Shaoxing⁴⁷ (Chapman); Ningbo First Hospital⁴⁸ (Ningbo); part of

the dataset used in Tianchi Arrhythmia Competition³ (Hefei); Shandong Provincial Hospital⁴⁹ (SPH); and SaMi-Trop⁵⁰. In addition, to create a hard benchmark that mirrors real-world cardiological practice—managing a wide range of cardiovascular conditions in patients with diverse health profiles—we harmonise labels from each dataset and merge all these sources into a new, comprehensive dataset named *Cardio-Learning*. This dataset includes 2.4 million ECGs from four countries, each labelled with one or more conditions from a comprehensive, highly imbalanced set of 164 conditions. After fine-tuning, which requires only a small fraction of pre-training time, HuBERT-ECG is evaluated on hold-out test sets of 17 datasets (16 individual + Cardio-Learning). The obtained performance is then compared with that achieved by training the same model from scratch, i.e., using randomly initialised weights. This comparison is common practice in self-supervised learning literature and reveals how self-supervised pre-training enhances both learning efficiency and generalisation. Where applicable, we also benchmark against existing state-of-the-art methods. To ensure consistent evaluation, address label distribution disparities, and employ a robust, threshold-independent metric, we report model performance using the Area Under the Receiver Operating Characteristic Curve (AUROC). The fine-tuning procedure is detailed in “Methods” - *Supervised Fine-tuning*. The following results, presented across diverse application scenarios, are meant to demonstrate the robustness and transferability of HuBERT-ECG’s self-supervised ECG representations.

HuBERT-ECG is efficient on small-sampled datasets. In diagnostic scenarios with small datasets, such as PTB (515 instances, 17 conditions), CPSC (6,878 ECGs, 9 conditions) and CPSC-Extra (3,453 ECGs, 52 conditions), HuBERT-ECG consistently shows solid performance and exhibits distinct macro-trends. On PTB (Fig. 3a), the SMALL and BASE configurations perform comparably well when fine-tuned, achieving AUROCs of 0.843 ± 0.029 and 0.848 ± 0.03 , respectively.

On CPSC (Fig. 3b), the SMALL model size achieves an AUROC of 0.945 ± 0.002 and is slightly surpassed by the BASE (0.949 ± 0.002) and LARGE (0.949 ± 0.006) models. On this dataset we outperform contrastive learning works (e.g., Liu et al.^{34,35}, Tian et al.³⁹ and Zhang et al.³⁷ achieved AUROCs of 0.932, 0.915, 0.910, and 0.920, respectively) while remaining competitive with models specifically pre-trained and optimised for arrhythmia detection only (e.g., Na et al.⁵¹ with 0.980).

On CPSC-Extra (Fig. 3c), a more challenging dataset, HuBERT-ECG SMALL performs best, reaching an AUROC of 0.894 ± 0.076 . This result illustrates the advantages of smaller architectures in such intricate, low-data contexts, where larger model configurations might overfit and perform less effectively (BASE 0.756 ± 0.02 ; LARGE 0.774 ± 0.023). Noteworthy, fine-tuned models also provide

³ <https://tianchi.aliyun.com/competition/entrance/231754/introduction>

0.03-0.058 AUROC improvements over their randomly initialised counterparts, with the less relevant exception of the BASE configuration on CPSC-Extra, demonstrating the clear benefit of pre-training even on smaller datasets.

HuBERT-ECG remains efficient in increasingly difficult contexts. Diagnostic complexity extends beyond simple metrics like dataset size or the number of classes. It is significantly influenced by the inherent subtlety and difficulty of identifying certain conditions, low per-class sample sizes, and highly skewed label distributions. These factors are well-exemplified by the Georgia and Chapman datasets, each featuring over 10,000 ECGs labelled with 65 and 52 conditions unevenly distributed. In these complex diagnostic scenarios, fine-tuning HuBERT-ECG consistently leads to better performance than training from scratch, yielding improvements of 0.065-0.078 on Georgia and 0.039-0.061 on Chapman.

HuBERT-ECG performance on downstream datasets varying weights initializations and model size

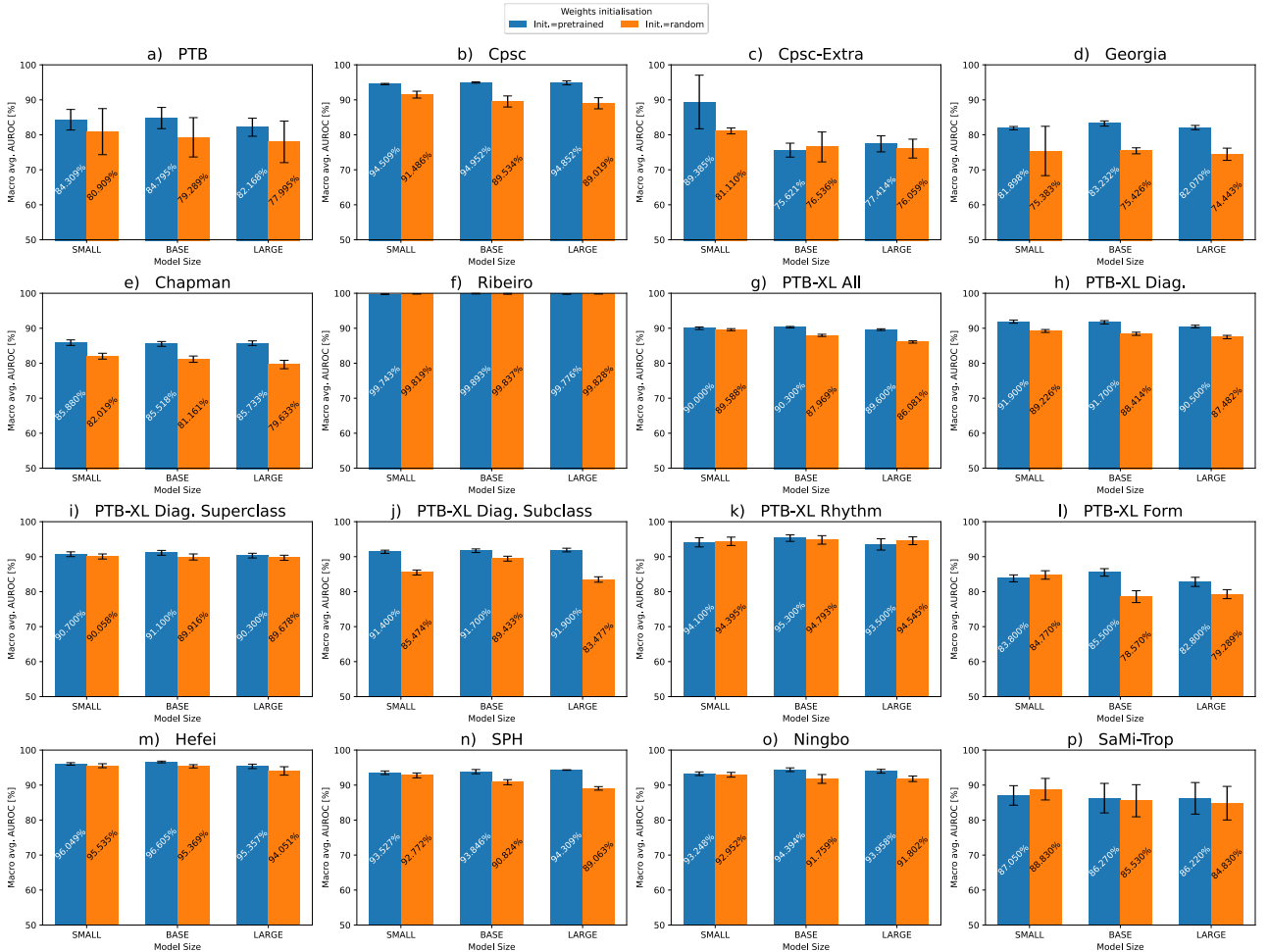


Figure 12. HuBERT-ECG macro-averaged performance across 16 datasets varying weights initialization and model size.

On Georgia, HuBERT-ECG BASE reaches an AUROC of 0.832 ± 0.007 (Fig. 3d). For the sake of comparison, Tian et al.³⁹ and Zhang et al.³⁷ facilitated the learning problem by limiting the diagnostic scope to 20 and 46 classes, reaching AUROCs of 0.900 and 0.866. When we fine-tune our BASE

configuration on these exact subsets, we substantially match the reported AUROCs: 0.895 ± 0.079 on the 20-class problem and 0.858 ± 0.093 on the 46-condition one. In contrast, when directly evaluated on the same condition subsets without retraining, despite being fine-tuned to detect all 65 diagnoses simultaneously, HuBERT-ECG BASE reaches competitive (0.853 ± 0.003) and matching (0.855 ± 0.093) AUROC scores.

On Chapman (**Fig. 3e**), all model configurations perform similarly, with the SMALL one topping (0.859 ± 0.008). To address the label distribution skewness, Chapman’s authors suggested grouping conditions into four macro-categories, highly simplifying the diagnostic task. Following this approach, Liu et al.³⁴ and Kiyasseh et al.³² reported AUROCs of 0.997 and 0.932, respectively. When our BASE configuration is fine-tuned under the same grouping, it matches the state-of-the-art with an AUROC of 0.9945 ± 0.0009 . Notably, when we directly test all model configurations on these four broader classes—despite being fine-tuned to recognise the original 52 co-occurring conditions—they achieve highly competitive AUROC values (SMALL 0.975 ± 0.002 ; BASE 0.975 ± 0.009 ; LARGE 0.984 ± 0.001) without marked performance loss. This strong performance underscores HUBERT-ECG’s robustness across varying levels of diagnostic granularity, even without simplifications of the underlying problem during training.

HuBERT-ECG enables accurate diagnostics in large-scale, complicated scenarios. Trained on extensive datasets, diagnostic foundation models are expected to perform robustly in large-scale, complicated scenarios, where they should identify a disease among many others, even accounting for significant inter-patient variations that can substantially affect diagnostic accuracy. To explore HuBERT-ECG’s transferability to these scenarios, we leverage large datasets with few to tens of possible conditions including Ribeiro, PTB-XL, Hefei, SPH and Ningbo.

On Ribeiro, a large-scale dataset (2.3 million ECGs, 6 conditions), every HuBERT-ECG configuration performs excellently regardless of weights initialization. The BASE model performs best when fine-tuned (0.9989) and surpasses the trained-from-scratch counterpart (0.9984) whereas the SMALL and LARGE configurations perform similarly (**Fig. 3f**). Table 2, which includes only the best model configuration in terms of AUROC alongside existing models for conciseness, shows that HuBERT-ECG BASE outperforms the model from Ribeiro et al.⁴² in terms of sensitivity, specificity, AUROC and area under precision and recall curve (AUPRC). We also surpass ECGFounder⁴⁰ on all these metrics. The full comparison including all model sizes, which provide comparable or slightly higher performance on other metrics, is provided in Supplementary Table 7.

The PTB-XL⁴⁵ dataset comprises 21,837 ECGs annotated with 71 possible labels covering diagnostic, form and rhythm statements. The authors state that 44 diagnostic conditions can be aggregated into 23 diagnostic subclasses and 5 broader diagnostic superclasses, while there are 12 possible rhythm

statements and 19 different form abnormalities. Therefore, one can evaluate a model on 6 different datasets, referred to as PTB-XL *All*, *Form*, *Rhythm*, *Diagnostic (Diag.)*, *Diag. Subclass*, *Diag. Superclass*, that differ from each other in the reported labels. On these sources, fine-tuned models surpass randomly initialised counterparts for almost each model configuration (**Fig. 3g-l**). Particularly, HuBERT-ECG BASE achieves an AUROC greater than 0.900 on 5 out of 6 datasets. An extensive comparison with works from the literature is presented in Supplementary Table 8. Many of these studies focus only on some PTB-XL class subsets (e.g., *All*, *Superclass*), limiting possible comparisons, and most of them leverage models specifically optimised for this ECG benchmark. In this context, the dataset-agnostically pre-trained HuBERT-ECG proves competitive with state-of-the-art works tailored on PTB-XL⁵² and surpasses other solutions^{34,35,37}. Unlike prior approaches such as KED³⁹, ECGFounder⁴⁰ and ECGFM³⁷—which restricted their scope to simplified subsets of PTB-XL *All* (46 diagnostic classes for KED and 30 for the others) and reported AUROC scores of 0.858, 0.924 and 0.917 respectively—our models demonstrate markedly superior performance, even under more challenging conditions. For fair benchmarking, we fine-tune our BASE model following the same configurations of these methods. Under identical conditions, our model reaches superior AUROCs: 0.911 ± 0.003 on the 46-condition problem and 0.921 ± 0.009 on the 30-class one. Differently, when directly evaluated on these class subsets, despite being fine-tuned for the full 71-class learning problem, HuBERT-ECG models achieve the following AUROCs: 0.900 (SMALL), 0.904 (BASE), and 0.901 (LARGE) on KED’s 46-class subset; 0.918 (SMALL), 0.926 (BASE), and 0.923 (LARGE) on ECGFounder and ECGFM’s 30-class subset. This collection of robust and superior performance, even under more challenging learning conditions, highlights the broader generalisability of HuBERT-ECG’s self-supervised representations compared to the those learnt by existing supervised or contrastive methods.

On Hefei (more than 20k samples, 29 conditions), all fine-tuned models surpass the respective randomly initialised counterparts, with the BASE configuration performing best with an AUROC of 0.9661 ± 0.0025 (Fig. 3m). On SPH⁴⁹, a collection of 25,770 ECGs annotated with 44 diagnostic statements, all our models achieve comparable results, with the LARGE model size reaching the highest AUROC (0.943) (Fig. 3n). To the best of our knowledge, we are the first to address this dataset in its full complexity.

Finally, on Ningbo⁴⁸ (34,905 ECGs, 76 co-existing conditions), fine-tuned models generally outperform the trained-from-scratch counterparts, with the BASE size achieving the highest AUROC of 0.9439 ± 0.0053 (Fig. 3o).

HuBERT-ECG can be a good predictor of future cardiovascular events. Predicting future CVEs is an exceptionally challenging task, especially given the scarcity of publicly available datasets. In this

context, the SaMi-Trop⁵⁰ dataset offers an opportunity to assess HuBERT-ECG’s predictive capabilities. It includes 1632 ECGs from patients with Chagas disease monitored during a 2-year follow-up period, 268 of which are normal recordings and 104 are marked as “death events”. On this dataset, the BASE and LARGE model configurations benefit more than the SMALL one from pre-trained weights. In fact, although competitive, the fine-tuned SMALL version reaches an AUROC of 0.8705 ± 0.0279 while the randomly initialised counterpart achieves a score of 0.8883 ± 0.0308 . In contrast, the BASE (0.8627 ± 0.0422) and LARGE (0.8622 ± 0.0451) configurations perform marginally better than their trained-from-scratch counterparts. We attribute this behaviour to the low number of events within SaMi-Trop, where a single misclassification can result in significant changes. For comparison purposes, we benchmark against the state-of-the-art work by Ferreira et al.⁵³. While we use only the raw ECG traces, they used a combination of handcrafted ECG features, sociodemographic variables and self-reported symptoms to fit a Random Forest estimator, and evaluate their model in terms of G_{mean} score, where $G_{mean} = \sqrt{sensitivity \cdot specificity}$. Fixing the same classification threshold, our best model surpasses theirs, reaching a G_{mean} of 0.78 against one of 0.77, while, in terms of AUROC, no comparison is possible due to the lack of predictions and code.

HuBERT-ECG navigates extremely complex scenarios. In real-world clinical practice, cardiologists are expected to diagnose a wide spectrum of heart-related conditions, adapting to the nuances of each patient’s health profile, regardless of how frequently a condition appears. To emulate this complexity, we employ the newly assembled, comprehensive *Cardio-Learning* dataset, which contains more than 2.4 million ECGs and 164 overlapping conditions, where the ECG is either the primary diagnostic tool, a non-primary supportive diagnostic tool, or used to estimate the risk of future CVEs. The fine-tuned HuBERT-ECG BASE provides a macro-averaged AUROC of 0.8517 ± 0.0193 , while HuBERT-ECG SMALL and LARGE follow with scores of 0.8453 ± 0.0058 and 0.8403 ± 0.0378 , respectively (**Fig. 4a-c**). When the conditions are grouped according to the ECG diagnostic role—primary diagnostic tool or informative support for diagnosis—the BASE configuration achieves averaged AUROCs of 0.84079 ± 0.019 and 0.8777 ± 0.0197 , respectively. Most interestingly, without any additional recordings and by leveraging inter-relationship among conditions, HuBERT-ECG raises AUROC values in predicting future deaths for SaMi-Trop patients to 0.9096 ± 0.0782 , significantly surpassing the best model presented for this task in the previous paragraph.

Benchmarking HuBERT-ECG against reported human performance. To contextualize the clinical relevance of HuBERT-ECG, we retrospectively compare the performance of its fine-tuned version on the Ribeiro dataset with that of human readers reported in Ribeiro et al.⁴². That study involved two

4th-year cardiology residents, two 3rd-year emergency residents, and two 5th-year medical students, whose sensitivity and specificity values were averaged across the six diagnostic categories.

Using the same dataset and label set, and the balanced accuracy as aggregation metric, we find that always at least two fine-tuned HuBERT-ECG configurations outperform medical residents/students on all the six diagnostic categories, and all configurations always surpass at least two groups of residents/students (see Table 3). However, it is important to note that this is not a direct head-to-head comparison, and no statistical testing can be performed due to the unavailability of individual predictions from the human annotators. Nonetheless, the results suggest that HuBERT-ECG, when fine-tuned, is capable of matching or surpassing human-level performance on key diagnostic tasks, highlighting its potential in downstream clinical applications.

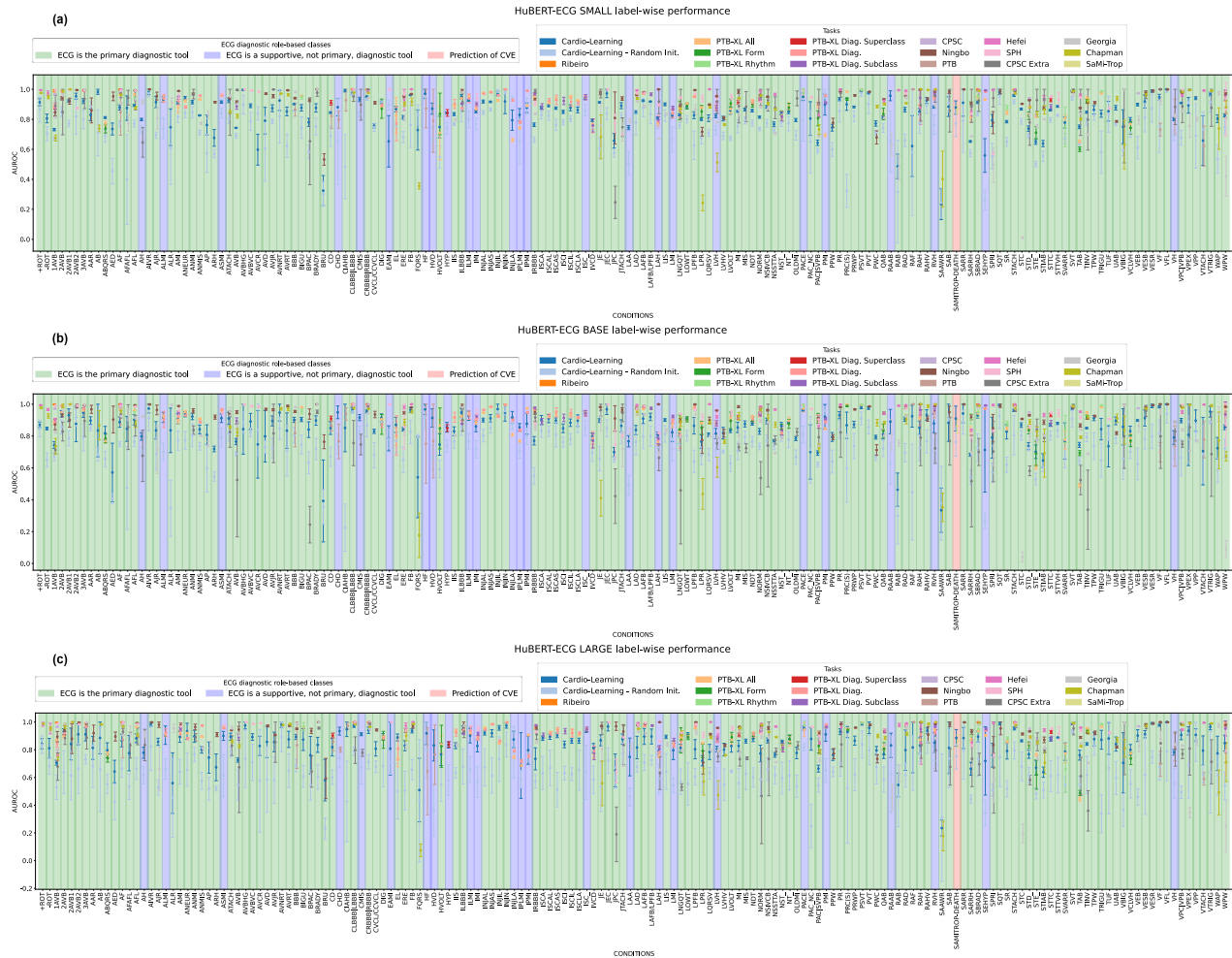


Figure 13. HuBERT-ECG label-wise performance for (a) SMALL, (b) BASE, and (c) LARGE configurations across all datasets, including Cardio-Learning. Label backgrounds are coloured based on the diagnostic role of the ECG (Zoom in for better view)

Fine-tuning HuBERT-ECG requires less compute time than training from scratch. While acquiring sufficient labelled data is already a known bottleneck, compute time represents an equally critical constraint in deep learning workflows, particularly in domains where large datasets and

complex models necessitate resource-intensive experimentation, hyperparameter optimization, and ablation studies. As demonstrated in Table 4, fine-tuning the pre-trained HuBERT-ECG is markedly more efficient across most scenarios. The aforementioned steps, not included in the presented results, are already performed during the self-supervised pre-training of HuBERT-ECG and can therefore be skipped by future users, offering a clear advantage over workflows that require developing models entirely from the ground up.

Discussion

In this paper, we introduce HuBERT-ECG, a novel foundation model for comprehensive ECG analysis available in three scalable configurations (SMALL, BASE, and LARGE) to address diverse clinical needs through a targeted fine-tuning. Pre-trained via self-supervision on an extensive dataset of 9.1 million ECGs from four countries and diverse populations, HuBERT-ECG demonstrates strong performance in identifying 164 cardiovascular conditions. This is rigorously validated across 16 individual datasets and their aggregation (Cardio-Learning), encompassing small to very large ECG collections, annotated with few to many overlapping conditions. Across all evaluations, the three model configurations consistently achieve AUROCs above 0.9 for an average of 136, 138, and 135 conditions and surpass 0.95 AUROC for an average of 95, 94, and 94 conditions, respectively (Fig. 5a). These results underscore the generalisability of HuBERT-ECG’s self-supervised representations to highly different clinical contexts. Looking ahead, our performance analysis aims not only to shed light on the data requirements for effective fine-tuning but also to guide performance optimisation on new ECG-based datasets. In diagnostic contexts with small-sampled datasets, which, requiring limited collection efforts, are common, HuBERT-ECG demonstrates both efficiency and accuracy, even when the number of conditions outpace the number of examples, often resulting in highly imbalanced label distributions. In these contexts, the SMALL and BASE configurations are preferable, demonstrating superior generalisation capabilities, whereas the LARGE configuration is less recommended. As the number of patients, ECGs and conditions increases, simulating more realistic diagnostic environments, HuBERT-ECG delivers remarkable results regardless of diagnostic granularity, albeit requiring adequate data collection efforts for fine-tuning. Furthermore, scaling up model size proves advantageous on larger sources, although smaller configurations remain competitive, thus providing desirable flexibility for complicated diagnostic fine-tuning. Notably, on Ribeiro, the largest individual dataset we use, HuBERT-ECG showcases the capacity to achieve excellent precision-sensitivity and specificity-sensitivity trade-offs (Table 2), matching or surpassing reported human performance (Table 3). Although Fig. 3f shows that the benefits of pre-training are marginal at the extremely high performance levels obtained from such large datasets with only a few

conditions, this does not diminish the value of our approach. In fact, such extensive data collection and labelling efforts—combined with only six conditions—are uncommon in real-world settings. Nonetheless, when benchmarking our performance against prior studies, we demonstrate our straightforward fine-tuning consistently leads to high performance, frequently approaching or surpassing that of models specifically developed for individual datasets or condition groups (e.g., arrhythmias). In addition, aiming at learning efficiency, we show that fine-tuning the pre-trained HuBERT-ECG generally requires less compute time than training from scratch, which also demands costly hyper-parameter tuning and ablation studies. Ultimately, when comparing with other methods proposed as foundation models, HuBERT-ECG showcases matching or stronger performance even under more challenging learning conditions. These findings underscore the high generalisability and utility of its self-supervised ECG representations, especially when data is overall scarce, certain classes are underrepresented, or when compute time is a considerable constraint.

Furthermore, HuBERT-ECG exhibits robust performance in highly complex clinical scenarios exemplified by Cardio-Learning, which aggregates ECGs from diverse sources and countries without task-specific separation and creates a realistic representation of heterogeneous real-world data. In this challenging context, HuBERT-ECG effectively identifies conditions for which the ECG serves as either a primary or informative support signal for diagnosis. Crucially, fine-tuning on Cardio-Learning enables HuBERT-ECG to leverage important inter-relationships among conditions, leading to substantial improvements in AUROC for label prediction—even for conditions that appear only in single subsets. For example, when fine-tuning the BASE model on PTB, the AUROCs for Coronary Heart Disease (CHD) and Heart Failure (HF) are 0.7656 and 0.7514, respectively. Remarkably, when fine-tuned on Cardio-Learning, these scores increase significantly to 0.9488 ± 0.0399 for CHD and 0.9674 ± 0.0566 for HF (**Fig. 5b**).

Similarly, in the highly challenging task of predicting future mortality events, HuBERT-ECG achieves mean AUROCs of 0.8784, 0.9096, and 0.8857 for the SMALL, BASE, and LARGE configurations, respectively, with average improvements of 1.31, 4.69, and 2.35 percentage points (Fig. 5c). These findings strongly suggest that training a model to recognize a broad spectrum of ECG-derived diagnostic labels can significantly enhance its performance in complex prediction tasks, including those where the ECG plays a non-primary but potentially informative role. Importantly, although the model is trained on labels from ECG readings, its rich representations can be further fine-tuned using high-quality clinical endpoints (e.g. electronic health records, imaging and/or genomic data), potentially unlocking even greater performance on tasks beyond traditional ECG applications.

Another relevant finding from HuBERT-ECG’s label-wise performance on Cardio-Learning is its low prediction variance (Fig. 4). This suggests that merged labels within Cardio-Learning refer to

consistent concepts, thus (1) excluding major label inconsistencies, and (2) supporting both our label harmonisation process and HuBERT-ECG's robustness to label noise. Furthermore, a visual inspection of the fine-tuned HuBERT-ECG attention mechanism confirms its ability to attend to clinically relevant ECG segments rather than arbitrary noise (see Supplementary Information – Sec. 3). While HuBERT-ECG has been rigorously tested across diverse cardiovascular conditions, our findings suggest its considerable potential to extend its application toward more personalised clinical insights. These include predicting patient response to therapies or identifying non-cardiac conditions through ECG analysis, paving the way for truly personalised cardiovascular care. For example, in the management of heart failure, HuBERT-ECG could be fine-tuned to predict patient responsiveness to resynchronisation therapy, thereby optimising patient selection criteria⁵⁴. This capability would facilitate a more tailored therapeutic approach, aimed not only at predicting adverse events but also at enhancing outcomes, reducing hospitalisations, and improving quality of life. Such advancements could fundamentally transform cardiovascular precision medicine, with the use of foundation ECG models for helping clinicians tailor therapies with greater precision⁵⁵.

Despite the highly encouraging results, our approach has certain limitations. First, although the data collection spans more countries than in previous studies, it does not include large and highly populated regions (e.g., Africa and India), which we aim to incorporate in future versions. Second, the scarcity of accessible ECG datasets for predicting future CVEs partially limits our ability to fully assess the HuBERT-ECG's potential in this still underexplored domain. Broader access to suitable datasets for testing the model in this challenging field is desirable to accelerate research. Third, direct and fair comparisons with previous systems, both deep-learning-based and rule-based, remain challenging. Many existing works focus on limited subsets of available datasets (often reducing the actual number of conditions), do not open-source their implementations or evaluation datasets, or do not follow standard practices in metric computation, hampering reproducibility and equitable benchmarking. Lastly, we observe that while some conditions are better classified in Cardio-Learning than in their original datasets, others show lower performance. This discrepancy may stem from erroneous labels or incomplete assignments. For example, in PTB-XL, a major label source, key terms in reports are extracted to retrospectively infer the physician's diagnostic confidence, with labels assigned even at low-confidence levels. Conversely, other datasets may rely on more robust labelling protocols, potentially resulting in identical ECGs from different sources receiving different diagnoses. Additionally, less experienced physicians may overlook certain conditions, while overly confident ones might exclude or place less emphasis on others, as also found by Li et al.⁴⁰. These inaccuracies can reduce the data efficiency, requiring more data to compensate for their effects.

Methods

Data and Preprocessing

While many ECG-focused studies addressing general clinical questions primarily rely on datasets from the Physionet Challenges^{56,57}, overlooking other substantial and valid sources, HuBERT-ECG self-supervised pre-training leverages both public and access-on-demand 12-lead ECG datasets. These include the labelled and unlabelled partitions from Ribeiro⁴² (except the predefined, publicly available test-set), CPSC and CPSC-Extra⁴³, PTB⁴⁴ and PTB-XL⁴⁵, the publicly available partition from Georgia⁴⁶, Chapman-Shaoxing⁴⁷ and Ningbo First Hospital⁴⁸, the partition from the Tianchi Arrhythmia Competition⁴, Shandong Provincial Hospital⁴⁹, and MIMIC-IV ECG⁵⁸. For supervised fine-tuning and evaluation we reuse all the aforementioned sources and include the new SaMi-Trop⁵⁰ dataset. We instead exclude the unlabelled partition of Ribeiro and MIMIC-IV for this phase. Although some datasets appear during both self-supervised pre-training and supervised fine-tuning, this is a widely accepted practice in developing foundation models that undergo a self-supervised pre-training. Notable examples include those by Caron et al.^{19,59} in computer vision, Hsu et al.²² and Baevski et al.²¹ in speech representation learning, Azizi et al.⁶⁰, Pai et al.⁶¹ and Haghighi et al.⁶² in various types of medical image analyses, Ritcher et al.⁶³ in cell genomics, and ultimately Mehary & Strodtloff³⁶ and Tang et al.⁶⁴ in the ECG domain. In addition, a recent survey⁶⁵ on the impact of self-supervised pre-training for diagnostic tasks involving many different types of medical data, points out that pre-training with self-supervision and fine-tuning a model on the same data brings no tangible benefit. This is due to the lack of any information leakage, as no clinical/downstream labels or task-specific signal is exposed during the self-supervised pre-training. In our case, pre-training target variables are also pseudo-labels, discovered in an unsupervised manner using feature descriptors (MFCCs and latent representations) that do not carry any disease-specific, clinical, or dataset-specific information. For self-supervised pre-training, all labels and clinical annotations are removed from the pre-training datasets, resulting in an unlabelled dataset of 9.1 million ECGs (**Fig. 2a–Pretrain**). The effectiveness of the data selection in capturing the diversity of ECG signals is illustrated through a UMAP⁶⁶ projection of the ECG embeddings (**Fig. 1e**). To assess HuBERT-ECG’s downstream utility on clinical tasks and datasets of varying complexity, we fine-tune it on each collected dataset⁵ and test it on hold-out partitions of such data sources. Additionally, we create a challenging new dataset, which we name *Cardio-Learning*, by merging all the above sources into a new one comprising over 2.4 million ECGs with 164 potentially co-occurring conditions. These conditions can be grouped into three categories based on the diagnostic role of the ECG, as shown in **Fig. 1f**. During the creation of

⁴ <https://tianchi.aliyun.com/competition/entrance/231754/introduction>

⁵ PTB-XL generates 6 different datasets that differ from each other in the presented conditions.

Cardio-Learning, to ensure label consistency across different sources, diagnostic labels are harmonised independently by three expert cardiologists. First, diagnostic labels that appear under slightly varying nomenclature are consolidated using unique terms. For example, all instances of atrial fibrillation are uniformly labelled as “AF” instead of “AFIB” or “AF_” (see Supplementary Table 1). Second, diagnostic concepts that appear under multiple synonyms are standardised to a single diagnostic category. For instance, following established practices^{56,57}, ECGs labelled as premature atrial contraction (PAC) or supraventricular premature beats (SVPB) are standardised to PAC|SVPB, as both conditions refer to the same concept. Third, hierarchically broader diagnostic categories are assigned to ECGs already labelled with more specific ones (e.g., Left Bundle Branch Block ECGs also receives the Bundle Branch Block label). However, we avoid two actions: (1) we do not infer more specific conditions from more general labels unless explicitly confirmed in the original annotation (e.g., Bundle Branch Block ECGs do not receive the Left Bundle Branch Block label); (2) if a constituent dataset includes both a more specific and a more general condition, but explicitly does not assign the general condition to ECGs labelled with the specific one, we respect this distinction and do not add it. Remarkably, as shown in **Figs. 1b and 1d**, the ECGs used in this study were obtained from patients exhibiting a broad age distribution and diverse geographical origins spanning four countries. To the best of our knowledge, this constitutes one of the largest and most heterogenous dataset ever assembled to date in terms of the number of conditions, demographics, and geographic origin of the individuals. For pre-processing, consistent with the approach of Natarajan et al.⁶⁷, a finite impulse response bandpass filter is applied to exclude frequencies outside the [0.05, 47] Hz range, which is reported to encompass the dominant components of P waves, T waves and QRS complexes⁶⁸. Subsequently, we investigate the impact of the sampling rate, which has no standard value and regulates the degree of dilution of the information content, on both the upstream and downstream performance. Our experimental findings indicate that resampling the ECGs at 100 Hz provides a suitable trade-off between downstream performance and training time, while preserving all the relevant physiological information content according to the Nyquist-Shannon theorem (Supplementary Information–Sec. 1.2). Following this, the ECG signal is rescaled to the [-1, 1] range, mirroring the procedure employed by Natarajan et al.⁶⁷. Finally, unlike other works, we use 5-second 12-lead ECGs instead of 10-second recordings, hence halving memory occupation and accelerating both training and inference. In addition, the selected temporal and spectral parameters align with those derived in Mehari & Strodthoff⁶⁹.

HuBERT-ECG Architecture and Theoretical Framework

A schematic representation of the HuBERT-ECG architecture, its pre-training phases and its fine-tuning are shown in **Fig. 2b**.

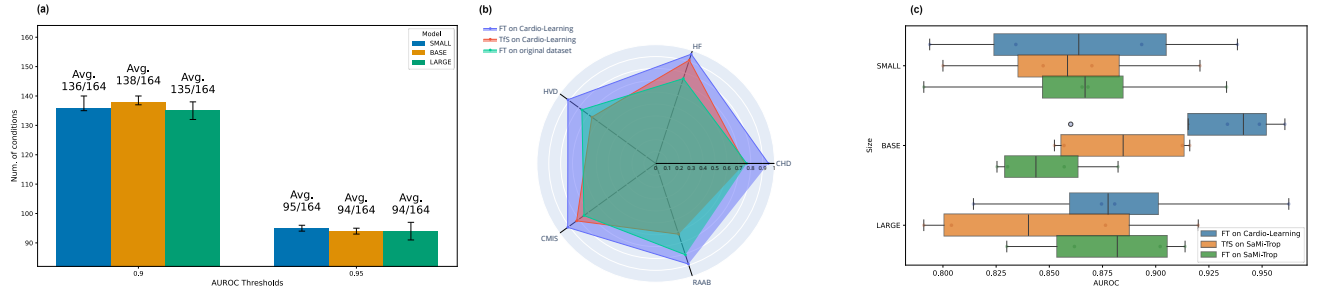


Figure 14. HuBERT-ECG’s performance overview with focus on diagnostic accuracy for conditions for which the ECG is used as an informative support of for predicting future CVEs. (a) average number of conditions, with min-max ranges, classified with at least 0.90 and 0.95 AUROC by HuBERT-ECG model configurations in at least one downstream dataset. (b) HuBERT-ECG averaged AUROC performance for conditions for which the ECG is an informative support for diagnosis and that appear only in one downstream dataset when the model is (1) fine-tuned on Cardio-Learning, (2) trained from scratch on Cardio-Learning, and (3) fine-tuned on the original downstream dataset. “FT” means fine-tuning; “TfS” means training from scratch. Conditions abbreviations: “HF” means Heart Failure; “CHD” means Coronary Heart Disease; “RAAB” means Right Atrial Abnormality; “CMIS” means Chronic Myocardial Ischemia; “HVD” means Heart Valve Disorder. (c) HuBERT-ECG AUROC trend in the prediction of death events for SaMi-Trop patients when it is fine-tuned on Cardio-Learning, trained from scratch on Sami-Trop, and fine-tuned on SaMi-Trop

Self-supervised pre-training

For each pre-training iteration we reserve 90% of the unlabelled dataset for training and the remaining 10% for internal validation. We set a batch size of 448 instances and find empirically the optimal masking percentage $p = 33\%$ (see Supplementary Information-Sec. 2.1) The first iteration consists of 80k training steps, while the second iteration counts 770k steps. The third iteration, instead, consists of 362.5k and 422.5k steps for the SMALL and LARGE model sizes, respectively. We use Adam optimiser with $\beta = (0.9, 0.98)$, an initial weight decay of 0.01 and a dropout probability of 0.1 along with a learning rate scheduler that ramps up for the first 8% of the training steps and then decays linearly to zero. Peak learning rates are set to $5e-5/5e-5/2.5e-5$ for the BASE, SMALL and LARGE model configurations, respectively. In addition, we find benefits from exploiting a *dynamic regularisation* during pre-training (Supplementary Information–Sec. 1.3). This technique “penalises” the model by incrementally increasing its dropout probability and weight decay if performance does not improve on the internal validation set for *penalty-count* consecutive times. Otherwise, if performance improves on the internal validation set, with respect to the best validation loss or accuracy, the model is “rewarded” by reversing the effects of the last penalty, i.e., decreasing its dropout probability and weight decay. It is important to note that the initial weight decay and dropout probability are the minimum achievable values, while *penalty-count* emerges as a significant hyperparameter to be tuned according to the frequency with which validations are performed. During the first and third iterations, with randomly initialised models, we set the *penalty-count* to 4 and

perform an internal validation every 2500 steps. For the second iteration, the internal validation is performed every 5000 steps, while maintaining the same value of *penalty-count*.

Supervised fine-tuning

In order to assess the robustness of HuBERT-ECG’s self-supervised representations on clinically relevant datasets simulating real-world scenarios, we fine-tune the model on each labelled dataset considered in this study. These datasets differ greatly from each other in the number of instances (Fig. 1a), patients’ age distribution (Fig. 1d), and spectrum of possible conditions, each belonging to one of the three classes outlined in Fig. 1f. When training-validation-test splits are predefined, or when at least the test set is known and defined a-priori, as in the case of PTB-XL, SPH or Ribeiro, we adhere to the original split to allow fair comparisons with previous works. However, for datasets where such a split is not available or applicable (i.e., Hefei, Ningbo, Chapman, CPSC, Georgia), we first extract a fixed hold-out test set, containing 10% of the dataset instances, using a stratified sampling to ensure that all the dataset diagnostic classes are represented. We then perform a 4-fold cross-validation on the remaining instances to tune the hyperparameters of the four models, selecting the best candidate from each fold for the evaluation on the hold-out test set. Finally, we average performance metrics that the four best candidates obtained during inference. Instead, for datasets with an extremely low cardinality compared to the model size, such as PTB, CPSC-Extra and SaMi-Trop, we split the dataset four times in a stratified manner into four $\langle \text{training}, \text{test} \rangle$ folds. Then, for each fold, lacking a validation set, we skip any hyperparameter tuning and train four models until either a near-zero training error is reached or a minimum improvement delta is not achieved. We then run inference directly on the corresponding hold-out test set with the last model checkpoint. When fine-tuning on PTB, CPSC-Extra and SaMi-Trop, to prevent overfitting on these small datasets, we set a minimal early stopping’s patience of 2 to allow for transitory fluctuations in the training loss descent and set a the minimum improvement delta to 1×10^{-3} for both the loss function and evaluation metrics. Lastly, to construct the Cardio-Learning dataset, we merge the training, validation, and test sets from each constituent dataset to form the corresponding overall training, validation, and test partitions, ensuring the prevention of any data leakage between these partitions. We perform a 4-fold cross-validation even on Cardio-Learning. Prior to any fine-tuning, we analyse the label distribution of each dataset and exclude the instances labelled with conditions that occur only once. We do so because these conditions are either unlearnable or untestable, as their solitary instances cannot be included in both the training and test sets. However, if a condition occurs twice in the dataset, we assign one instance to the training set and one to the test set, allowing the model to attempt to learn the condition while providing a way to assess its generalisability within the limits of this setup. Eventually, when a condition occurs three times, we place one example per split. During each fine-tuning procedure, we

optimise a binary cross-entropy loss function but refer to the macro-averaged multi-label AUROC on the validation set for checkpointing and model selection. The model checkpoint that yields the highest AUROC is used for inference. To fine-tune HuBERT-ECG we follow a straightforward protocol: we attach atop the pre-trained model a randomly initialised linear layer and fine-tune all the weights of the resulting architecture, except those of the convolutional embedder. In contrast to Hsu et al.²², we do not use the *freezing-steps* hyperparameter, as our experiments do not demonstrate any efficacy in maintaining the encoder’s parameters fixed while training only the final linear layer. Drawing inspiration from Devlin et al.¹⁷, we reduce the batch size to 64 instances and the learning rate to 1e-5. Also, to gain finer control over the identification of a good model candidate, we validate our models every 50 or 500 steps, contingent on the dataset size, hence more frequently than during pre-training. Due to the high number of trainable parameters and the limited size of most datasets, HuBERT-ECG shows a tendency to overfitting but we observe no benefits from using either a strong dropout (up to 0.5) or a high weight decay (up to 0.1), nor from freezing some encoding layers. However, experiments with LayerDrop⁷⁴ ([0.0, 0.1, 0.15, 0.2]) indicate its potential to mitigate validation loss divergence and metrics degradation. In addition, we note improvements when using a time-aligned random crop as data augmentation technique, a strategy that we also replicate at test time when we take the most confident prediction among those made on multiple crops of the same instance. In summary, our findings suggest that the best performance is obtained when fine-tuning the entire Transformer encoder, zeroing the dropout probability, maintaining the weight decay at 0.01, and sweeping over the LayerDrop probability. We track all experimental runs using Weights and Biases⁶.

Data availability

The PTB, CPSC, CPSC-Extra, Georgia, Chapman, PTB-XL, and Ningbo datasets are publicly accessible through PhysioNet at the following URL: <https://physionet.org/content/challenge-2021/1.0.3/training/ningbo/#files-panel>.

The Hefei dataset used in this study can be downloaded from the Tianchi platform at: <https://tianchi.aliyun.com/competition/entrance/231754/information/>.

The SPH dataset is available through Figshare at the following Springer Nature repository: https://springernature.figshare.com/collections/A_large-scale_multi-label_12-lead_electrocardiogram_database_with_standardized_diagnostic_statements/5779802/1.

The SaMi-Trop dataset can be downloaded from Zenodo at: <https://zenodo.org/records/4905618>.

The Ribeiro dataset is accessible for scientific research upon request to the respective owner. However, its test set is publicly available from Zenodo at: <https://zenodo.org/records/3765780>.

⁶ <https://wandb.ai/site>

Code availability

The full pipeline utilised in this study is available at the following GitHub repository: <https://github.com/Edoar-do/HuBERT-ECG>. This repository includes: (1) code for data preprocessing, from raw data management to the creation of the train-validation-test partitions used in our research; (2) scripts for replicating the pre-training, fine-tuning and inference of every model developed across all datasets; and (3) code for reproducing our performance validation. Eventually, to both facilitate reproducibility and enable rapid implementation, we provide (4) pre-trained weights for all model configurations on Hugging Face⁷. (5) Access to model weights fine-tuned on Cardio-Learning is on-demand. This model is intended as a more disease-oriented baseline for further fine-tuning.

References

1. Smilowitz, N. R. & Berger, J. S. Perioperative Cardiovascular Risk Assessment and Management for Noncardiac Surgery. *JAMA* 324, 279 (2020).
2. van de Vegte, Y. J. et al. Genetic insights into resting heart rate and its role in cardiovascular disease. *Nat Commun* 14, 4646 (2023).
3. Friedman, P. A. The Electrocardiogram at 100 Years: History and Future. *Circulation* 149, 411–413 (2024).
4. Ahsan, M. M. & Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artif Intell Med* 128, 102289 (2022).
5. Kalmady, S. V. et al. Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. *NPJ Digit Med* 7, 133 (2024).
6. Sun, W. et al. ECG for high-throughput screening of multiple diseases: Proof-of-concept using multi-diagnosis deep learning from population-based datasets. Preprint at <https://arxiv.org/pdf/2210.06291>. (2022).
7. Goto, S. et al. Multinational Federated Learning Approach to Train ECG and Echocardiogram Models for Hypertrophic Cardiomyopathy Detection. *Circulation* 146, 755–769 (2022).
8. Kim, M. et al. Deep learning for predicting rehospitalization in acute heart failure: Model foundation and external validation. *ESC Heart Fail* (2024).
9. Musa, N. et al. A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram. *J Ambient Intell Humaniz Comput* 14, 9677–9750 (2023).
10. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* 616, 259–265 (2023).
11. Finlayson, S. G. et al. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine* 385, 283–286 (2021).
12. Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. Preprint at <https://arxiv.org/abs/2108.07258>. (2021).
13. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training. (2018).
14. Radford, A. et al. Language Models are Unsupervised Multitask Learners. OpenAI blog 1, 9 (2019).
15. Brown, T. B. et al. Language Models are Few-Shot Learners. *Adv Neural Inf Process Syst* (2020).

⁷ <https://huggingface.co/Edoardo-BS>

16. OpenAI et al. GPT-4 Technical Report. Preprint at <https://arxiv.org/abs/2303.08774>. (2023).
17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2019. 4171–4186 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019).
18. Raffel, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* (2020).
19. Caron, M. et al. Emerging Properties in Self-Supervised Vision Transformers. in Proceedings of the IEEE/CVF international conference on computer vision (2021).
20. He, K. et al. Masked Autoencoders Are Scalable Vision Learners. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 16000–16009 (2022).
21. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in neural information systems* 33, 12449–12460 (2020).
22. Hsu, W.-N. et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans Audio Speech Lang Process* 29, 3451–3460 (2021).
23. Schäfer, R. et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat Comput Sci* 4, 495–509 (2024).
24. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat Med* 30, 863–874 (2024).
25. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* 630, 181–188 (2024).
26. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat Med* 30, 2924–2935 (2024).
27. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat Commun* 14, 4542 (2023).
28. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat Biomed Eng* 6, 1399–1406 (2022).
29. Ma, J. et al. Segment anything in medical images. *Nat Commun* 15, 654 (2024).
30. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nat Med* 30, 1481–1488 (2024).
31. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med* (2024).
32. Kiyasseh, D., Tingting Zhu & David A. Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. in Proceedings of the International Conference on Machine Learning 5606–5615 (2021).
33. Oh, J., Chung, H., Kwon, J., Hong, D. & Choi, E. Lead-agnostic self-supervised learning for local and global representations of electrocardiograms. in Conference on Health, Inference and Learning 338–353 (2022).
34. Liu, W. et al. Dense lead contrast for self-supervised representation learning of multilead electrocardiograms. *Inf Sci (N Y)* 634, 189–205 (2023).
35. Liu, W. et al. A joint cross-dimensional contrastive learning framework for 12-lead ECGs and its heterogeneous deployment on SoC. *Comput Biol Med* 152, 106390 (2023).
36. Mehari, T. & Strodthoff, N. Self-supervised representation learning from 12-lead ECG data. *Comput Biol Med* 141, 105114 (2022).
37. Zhang, S. et al. ECGFM: A foundation model for ECG analysis trained on a multi-center million-ECG dataset. *Information Fusion* 124, 103363 (2025).
38. Vaid, A. et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digit Med* 6, 108 (2023).
39. Tian, Y. et al. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Rep Med* 5, 101875 (2024).

40. Li, J. et al. An Electrocardiogram Foundation Model Built on over 10 Million Recordings with External Evaluation across Multiple Domains. Preprint at <https://arxiv.org/abs/2410.04133>. (2025).
41. Xu, M. et al. HMM-Based Audio Keyword Generation. in 566–574 (2004).
42. Ribeiro, A. H. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 11, 1760 (2020).
43. Liu, F. et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. *J Med Imaging Health Inform* 8, 1368–1373 (2018).
44. Bousseljot, R., Kreiseler, D. & Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik/Biomedical Engineering* 317–318 (2009).
45. Wagner, P. et al. PTB-XL, a large publicly available electrocardiography dataset. *Sci Data* 7, 154 (2020).
46. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101, (2000).
47. Zheng, J. et al. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Sci Data* 7, 48 (2020).
48. Zheng, J. et al. Optimal Multi-Stage Arrhythmia Classification Approach. *Sci Rep* 10, 2898 (2020).
49. Liu, H. et al. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Sci Data* 9, 272 (2022).
50. Cardoso, C. S. et al. Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. *BMJ Open* 6, e011181 (2016).
51. Na, Y., Park, M., Tae, Y. & Joo, S. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. in *Proceedings of the International Conference on Learning Representations* (2024).
52. Strodthoff, N., Wagner, P., Schaeffter, T. & Samek, W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *IEEE J Biomed Health Inform* 25, 1519–1528 (2021).
53. Ferreira, A. M. et al. Two-year death prediction models among patients with Chagas Disease using machine learning-based methods. *PLoS Negl Trop Dis* 16, e0010356 (2022).
54. Wouters, P. C. et al. Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. *Eur Heart J* 44, 680–692 (2023).
55. Leclercq, C. et al. Wearables, telemedicine, and artificial intelligence in arrhythmias and heart failure: Proceedings of the European Society of Cardiology Cardiovascular Round Table. *EP Europace* 24, 1372–1383 (2022).
56. Perez Alday, E. A. et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiol Meas* 41, 124003 (2020).
57. Reyna, M. A. et al. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. in *2021 Computing in Cardiology (CinC) 1–4* (IEEE, 2021).
58. Gow, B. et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. (2023).
59. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).
60. Azizi, S. et al. Big Self-Supervised Models Advance Medical Image Classifications. in *Proceedings of the IEE/CVF international conference on computer vision* 3478–3488 (2021).
61. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat Mach Intell* 6, 354–367 (2024).
62. Haghighi, F., Taher, M. R. H., Gotway, M. B. & Liang, Ji. DiRA: Discriminative, Restorative, and Adversarial Learning for Self-supervised Medical Image Analysis. in *Proceedings of the IEE/CVF international conference on computer vision* 20824–20834 (2022).
63. Richter, T., Bahrami, M., Xia, Y., Fischer, D. S. & Theis, F. J. Delineating the effective use of self-supervised learning in single-cell genomics. *Nat Mach Intell* 7, 68–78 (2024).

64. Tang, J., Xia, T., Lu, Y., Mascolo, C. & Saeed, A. Electrocardiogram Report Generation and Question Answering via Retrieval-Augmented Self-Supervised Modeling. in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1–5 (IEEE, 2025).
65. VanBerlo, B., Hoey, J. & Wong, A. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. *BMC Med Imaging* 24, 79 (2024).
66. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* 3, 861 (2018).
67. Natarajan, A. et al. A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification. in *Computing in Cardiology* (2020).
68. Hong, S., Zhang, W., Sun, C., Zhou, Y. & Li, H. Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020. *Front Physiol* 12, (2022).
69. Mehari, T. & Strodthoff, N. Towards Quantitative Precision for ECG Analysis: Leveraging State Space Models, Self-Supervision and Patient Metadata. *IEEE J Biomed Health Inform* 27, 5326–5334 (2023).
70. Choi, S. et al. ECGBERT: Understanding Hidden Language of ECGs with Self-Supervised Representation Learning. Preprint at <https://arxiv.org/abs/2306.06340>. (2023).
71. Vaswani, A. et al. Attention Is All You Need. *Adv Neural Inf Process Syst* (2017).
72. Hu, R., Chen, J. & Zhou, L. Spatiotemporal self-supervised representation learning from multi-lead ECG signals. *Biomed Signal Process Control* 84, 104772 (2023).
73. Arthur, D. & Vassilvitskii, S. k-means++: the advantages of careful seeding. in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027–1035 (ACM-SIAM Symposium on Discrete Algorithms, 2007).
74. Fan, A., Grave, E. & Joulin, A. Reducing Transformer Depth on Demand with Structured Dropout. in *Proceedings of the International Conference on Learning Representations* (2019).

Acknowledgements

This work has been partly funded by 1) Regione Lombardia, Italy, through the initiative “Programme of measures for economic recovery: development of new cooperation agreements with universities for research, innovation and technology transfer”-DGR n. XI/4445/2021; 2) European Union-Next Generation EU, through the Italian Ministry of Research PRIN 2022, project n. 2022A49KR3 “QT-SEED Quality-of-life Technological and Societal Exploitation of ECG Diagnostics”.

We also thank the authors of the CODE study⁴² for giving us access to the Ribeiro dataset.

Author Contribution

Edoardo Coppola: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing-original draft, Writing-review & editing. **Mattia Savardi:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Visualization, Writing-original draft, Writing-review & editing. **Mauro Massussi:** Conceptualization, Investigation, Writing-original draft, Writing-review & editing. **Marianna Adamo:** Investigation, Writing-review & editing. **Marco Metra:** Investigation, Writing-review & editing. **Alberto Signoroni:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing-original draft, Writing-review & editing.

Ethics statement

For the Ribeiro dataset the CODE Study was approved by the Research Ethics Committee of the Universidade Federal de Minas Gerais, protocol 49368496317.7.0000.5149.

		SMALL	BASE	LARGE
Convolutional embedder	strides	4, 2, 2, 2, 2		
	kernels width	10, 3, 3, 2, 2		
	channels	512		
Transformer encoder	layers	8	12	16
	internal dimension	512	768	960
	feed-forward dimension	2048	3072	3840
	attention heads	8	8	12
Embedding dimension W		256	256	512
Number of parameters		30M	93M	188M

Table 1 | HuBERT-ECG architecture summary.

Condition	Models	Sensitivity	Specificity	AUROC	AUPRC
AF	Ribeiro et al.	0.769	1.000	0.885	0.773
	Li et al.	1.000	0.996	0.999	N.A.
	HuBERT-ECG BASE	1.0000 (CI:0.9999– 1.0000)	1.0000 (CI:0.9999– 1.0000)	1.0000 (CI:0.9999– 1.0000)	1.0000 (CI:0.9999– 1.0000)
IAVB	Ribeiro et al.	0.929	0.995	0.962	0.807
	Li et al.	0.864	0.957	0.949	N.A.
	HuBERT-ECG BASE	0.7500 (CI:0.6071- 0.8929)	0.9981 (CI:0.9962- 1.0000)	0.9977 (CI:0.9961- 0.9994)	0.9560 (CI:0.9264- 0.9856)
CLBBB LBBB	Ribeiro et al.	1.000	1.000	1.000	1.000
	Li et al.	1.000	0.996	0.998	N.A.
	HuBERT-ECG BASE	0.9000 (CI:0.8333- 0.9667)	1.0000 (CI:0.9999– 1.0000)	1.0000 (CI:0.9999– 1.0000)	1.0000 (CI:0.9999– 1.0000)
CRBBB RBBB	Ribeiro et al.	1.000	0.995	0.997	0.895
	Li et al.	0.971	0.971	0.989	N.A.
	HuBERT-ECG BASE	0.9400	0.9943	0.9986	0.9740

		(CI:0.8529-1.0000)	(CI:0.9899-0.9987)	(CI:0.9974-0.9998)	(CI:0.9513-0.9967)
SBRAD	Ribeiro et al.	0.938	0.996	0.967	0.782
	Li et al.	1.000	0.955	0.967	N.A.
	HuBERT-ECG BASE	1.0000 (CI:0.9999–1.0000)	0.9919 (CI:0.9864-0.9975)	0.9980 (CI:0.9954-1.0000)	0.8860 (CI:0.7504-1.0000)
STACH	Ribeiro et al.	0.937	0.997	0.985	0.923
	Li et al.	0.974	0.970	0.989	N.A.
	HuBERT-ECG BASE	0.9730 (CI:0.9729–0.9731)	0.9950 (CI:0.9913-0.9987)	0.9990 (CI:0.9983-0.9997)	0.9640 (CI:0.9342-0.9938)
MACRO-AVERAGED	Ribeiro et al.	0.935	0.997	0.966	0.863
	Li et al.	0.9682	0.9742	0.9818	N.A.
	HuBERT-ECG BASE	0.9255 (CI:0.8943-0.9567)	0.9969 (CI:0.9954-0.9983)	0.9989 (CI:0.9983-0.9995)	0.9630 (CI:0.9403-0.9857)

Table 2 | Fine-tuned HuBERT-ECG BASE performance on Ribeiro benchmarked against that from Ribeiro et al.⁴² and Li et al.⁴⁰ according to multiple metrics. Sensitivity and Specificity are obtained setting the threshold to 0.5, as done by Ribeiro et al., for a fair comparison. Performance uncertainty is reported using 95% confidence intervals obtained via empirical bootstrapping using 1000 bootstrap iterations. Label abbreviations and corresponding diagnosis can be found in Supplementary Table 1. A full comparison of all HuBERT-ECG model sizes, which may perform better than the BASE one on certain conditions for certain metrics, is provided in Supplementary Table 7.

Condition	Predictor	Sensitivity	Specificity	Balanced Accuracy
AF	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.0000)	0.9754 (CI:0.9656-0.9853)	0.9877 (CI:0.9827-0.9926)
	HuBERT-ECG BASE	1.0000 (CI:0.999–1.000)	1.0000 (CI:0.9999–1.0000)	1.0000 (CI:0.9999–1.0000)
	HuBERT-ECG LARGE	1.0000 (CI:0.9999–1.000)	0.9748 (CI:0.9644–0.9853)	0.9874 (CI:0.9821-0.9926)
	Cardiology residents	0.7690	0.9960	0.8825
	Emergency residents	0.6150	0.9980	0.8065
	Medical students	0.9230	0.9890	0.9560
IAVB	HuBERT-ECG SMALL	0.9290 (CI:0.8580-1.0000)	0.9531 (CI:0.9387-0.9675)	0.9411 (CI:0.8983-0.9837)
	HuBERT-ECG BASE	0.7500 (CI:0.6071-0.8929)	0.9981 (CI:0.9962-1.0000)	0.8741 (CI:0.8016-0.9464)
	HuBERT-ECG LARGE	1.0000 (CI:0.9990–1.0000)	0.9485 (CI:0.9337–0.9637)	0.97425 (CI:0.9663-0.9818)
	Cardiology residents	0.6790	0.9970	0.838
	Emergency residents	0.8210	0.9840	0.9025
	Medical students	0.929	0.9790	0.9540
CLBBB LBBB	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.0000)	0.9940 (CI:0.9893-0.9987)	0.997 (CI:0.9946-0.9993)
	HuBERT-ECG BASE	0.9000 (CI:0.8333-0.9667)	1.0000 (CI:0.9999–1.0000)	0.9500 (CI:0.9166-0.9833)
	HuBERT-ECG LARGE	0.9330 (CI:0.8993–0.9667)	0.9950 (CI:0.9913–0.9987)	0.9640 (CI:0.9453-0.9827)
	Cardiology residents	0.9000	1.0000	0.9500
	Emergency residents	0.8670	0.9990	0.9330
	Medical students	0.9000	0.9970	0.9485
CRBBB RBBB	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.000)	0.9850 (CI:0.9776-0.9924)	0.9672 (CI:0.9887-0.9962)
	HuBERT-ECG BASE	0.9400 (CI:0.8529-1.0000)	0.9943 (CI:0.9899-0.9987)	0.9792 (CI:0.9214-0.9993)
	HuBERT-ECG LARGE	0.9710 (CI:0.9420–1.0000)	0.9874 (CI:0.9798–0.9950)	0.9825 (CI:0.9609-0.9975)
	Cardiology residents	0.9710	0.9940	0.8820
	Emergency residents	0.7650	0.9990	0.9685
	Medical students	0.9410	0.9960	0.9904
SBRAD	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.0000)	0.9808 (CI:0.9716-0.9901)	0.9960 (CI:0.9857-0.9950)
	HuBERT-ECG BASE	1.0000 (CI:0.9999–1.0000)	0.9919 (CI:0.9864-0.9975)	0.9990 (CI:0.9931-0.9987)
	HuBERT-ECG LARGE	1.0000 (CI:0.9999–1.0000)	0.9796 (CI:0.9704–0.9889)	0.9670 (CI:0.9851-0.9944)
	Cardiology residents	0.9380	0.9960	0.9355
	Emergency residents	0.8750	0.9960	0.8725
	Medical students	0.7500	0.9950	0.9773
STACH	HuBERT-ECG SMALL	0.9730 (CI:0.9729–0.9731)	0.9816 (CI:0.9722-0.9911)	0.9773 (CI:0.9725-0.9821)
	HuBERT-ECG BASE	0.9730 (CI:0.9729–0.9731)	0.9950 (CI:0.9913-0.9987)	0.9840 (CI:0.9821-0.9859)

	HuBERT-ECG LARGE	0.9730 (CI:0.9728–0.9732)	0.9829 (CI:0.9734–0.9924)	0.9790 (CI:0.9731-0.9828)
	Cardiology residents	0.8110	0.9990	0.9050
	Emergency residents	0.9460	0.9970	0.9715
	Medical students	0.8380	0.9960	0.9170

Table 3 | Fine-tuned HuBERT-ECG performance on Ribeiro benchmarked against two 4th year cardiology residents (Cardiology residents), two 3rd year emergency residents (Emergency residents), two 5th year medical students (Medical students) involved in the study by Ribeiro et al.⁴². Both residents' and students' performance are averages, as reported in the original study. Label abbreviations and corresponding diagnoses can be found in Supplementary Table 1.

Dataset	Model size		
	HuBERT-ECG SMALL (FT / TfS)	HuBERT-ECG BASE (FT / TfS)	HuBERT-ECG LARGE (FT / TfS)
PTB	3.3k / 3.5k	2.3k / 2.15k	1.8k / 1.5k
Cpsc	5.8k / 6.7k	4.8k / 4.4k	3.2k / 3.1k
Cpsc-Extra	9.2k / 8.15k	6.7k / 3.9k	5.3k / 2.8k
Georgia	5.2k / 8.13k	5.4k / 6.3k	2.8k / 4.8k
Chapman	5k / 9.22k	4.4k / 5.3k	3.3k / 3.5k
Ribeiro	55k / 60k	45k / 55k	50k / 40k
PTB-XL All	12.5k / 27k	9k / 19k	8.5k / 19k
PTB-XL Diag.	11k / 21k	9.5k / 24.5k	7.5k / 20k
PTB-XL Diag. Superclass	7k / 18.5k	6k / 15.5k	4.5k / 16.5k
PTB-XL Diag. Subclass	8k / 12.5k	6.5k / 24.5k	8k / 8k
PTB-XL Rhythm	7.5k / 19.5k	8.5k / 25.5k	5.5k / 20k
PTB-XL Form	6.5k / 17k	6.5k / 9k	5.5k / 9k
Hefei	10.1k / 18.2k	5.8k / 15.2k	5.2k / 13.7k
SPH	10.5k / 27k	5.5k / 14.5k	7.5k / 21k
Ningbo	15.8k / 35.5k	12.6k / 21.7k	10.1k / 20k
SaMi-Trop	0.9k / 0.8k	1.1k / 0.6k	0.8k / 0.5k
Cardio-Learning	30k / 30.1k	27.8k / 42.1k	32.5k / 27.4k

Table 4 | Comparison between fine-tuning (FT) and training-from-scratch time (TfS), on average, across datasets and model sizes. Time is expressed as number of training steps.

3. Use Case 3: The Dorian Gray Project – Linking Cardiovascular Disease and Cognitive decline

4.1 The DORIAN GRAY Project: Architectural Design for a Cardiovascular Digital Twin

The concept of the digital twin—a dynamic, virtual replica of a physical system—will be transforming healthcare by enabling real-time simulation, forecasting, and optimization of complex processes²⁵. Its application in healthcare, although much more recent, is rapidly gaining momentum. In a clinical context, a digital twin can be described as an *in silico* representation of an individual that integrates multimodal and continuously updated data to model health trajectories, predict outcomes, and support personalized therapeutic decisions. These virtual replicas evolve through a constant exchange of information with their real-world counterparts, creating an interconnected “cyber-physical” loop capable of adaptive learning and real-time feedback.

Despite its promise, translating digital twin technology into medicine faces substantial challenges. Clinical data are often heterogeneous, fragmented, and acquired irregularly; biological models of disease remain incomplete; and many determinants of health—particularly behavioral and social factors—are difficult to capture systematically. As a result, current healthcare digital twin initiatives vary widely. Some rely only on static or retrospective datasets to forecast future health states, functioning as passive predictive tools. Others aim for a more sophisticated active design, continuously incorporating real-time streams from wearables, mobile devices, or environmental sensors to dynamically refine predictions. The latter approach, although far more powerful, requires robust multimodal integration, advanced AI methodologies with short reaction times, and a coordinated clinical ecosystem capable of supporting continuous data exchange.

In this evolving landscape, the European Union has launched several initiatives to explore next-generation digital health tools. Among them, the Horizon Europe–funded project DORIAN GRAY (“Devising a Personalized Risk Stratification and Holistic Management for Prevention of Cognitive Impairment in Patients with Different Cardiovascular Phenotypes,” grant agreement 101156266) represents one of the first large-scale efforts to develop a clinically actionable digital twin for the prevention of cognitive decline in individuals with cardiovascular disease (CVD) or at high cardiovascular risk²⁶.

The project is grounded in the growing recognition that cardiovascular disorders and mild cognitive impairment (MCI) are deeply interconnected. Conditions such as heart failure accelerate the biological mechanisms leading to cognitive deterioration, while traditional cardiovascular risk factors—hypertension, diabetes, obesity, sedentary lifestyle—also predispose individuals to neurocognitive decline²⁷. The central ambition of DORIAN GRAY is therefore twofold: first, to elucidate the mechanistic pathways linking CVD and MCI; and second, to exploit these insights to build a personalised, AI-augmented digital ecosystem that can identify individuals at risk, intervene early, and promote resilience in ageing populations.

To achieve this vision, the project integrates traditional clinical variables with emerging digital biomarkers captured through real-world data (RWD). Genomic, metabolomic, proteomic, imaging, and lifestyle information from established datasets will be harmonised to build the foundational structure of the digital twin within the PORTRAIT platform. This core model will then be enriched by continuous data streams—heart rate, sleep metrics, physical activity, behavioural patterns—collected via smartwatches and mobile devices during the clinical studies. Such a combination of high-resolution biological profiling and real-time behavioural monitoring provides a uniquely comprehensive view of the factors that influence cognitive performance and its deterioration.

One of the most innovative components of DORIAN GRAY is the development of an avatar-based coaching exergaming (ABCE) system, implemented through the MIRROR app. This tool merges two synergistic functions: a digital cognitive-physical training environment (the “exergame”) and an AI-driven coaching system that shapes healthy behaviours. Each participant interacts with a customizable, neutral-shaped virtual avatar that guides them through physical-cognitive tasks, provides real-time feedback, and adapts difficulty based on performance. While avatar technologies have previously been explored in post-stroke rehabilitation, their application to cognitive impairment and dementia prevention remains largely unexplored. The ABCE therefore represents a pioneering model of multidomain intervention that simultaneously targets physical activity, cognitive stimulation, lifestyle change, and social engagement.

The coaching system extends well beyond traditional decision support. Using adaptive and persuasive strategies, it encourages individuals to adopt healthier routines in daily life—improving nutrition, increasing activity levels, enhancing sleep quality, and mitigating behaviours known to accelerate cognitive decline. Healthcare professionals supervise the personalised plan through a dedicated clinician dashboard, adjusting goals and monitoring progress. Importantly, relatives and friends are also incorporated into selected exergame tasks, transforming the intervention into a socially enriched experience, which has been shown to reinforce motivation and adherence.

The DORIAN GRAY framework will undergo rigorous evaluation through two clinical studies. First, a pilot randomized trial in patients with heart failure and MCI will assess feasibility, usability, and early signals of clinical impact in a population with limited therapeutic options. Subsequently, an implementation study will test adherence and real-world acceptability in individuals with MCI and cardiovascular risk factors. Together, these studies will generate evidence on the effectiveness of the multidomain digital intervention, previously validated in nondigital formats, now enhanced by AI-driven adaptability and continuous monitoring.

Beyond clinical validation, the project includes a comprehensive socio-economic, ethical, and legal assessment to inform the potential deployment of the solution at scale. Partner organisations—including experts in exploitation and innovation management—will support the definition of Key Exploitable Results (KERs) and the design of pathways for sustained adoption beyond the project’s duration. This strategy ensures alignment between scientific development, regulatory considerations, communication, and real-world implementation.

Ultimately, DORIAN GRAY showcases how digital twins, multimodal AI, and real-world data can converge into an innovative and clinically meaningful system for personalised prevention. By identifying early “fingerprints” of cognitive decline, modelling individual trajectories, and delivering tailored interventions through an engaging avatar-based platform, the project embodies a new paradigm for the management of non-communicable diseases. Its integrative approach lays the foundation for future digital twin applications across multiple conditions, marking a significant advance toward the European vision of personalised, predictive, participatory, and preventive healthcare.

Global Spotlights

The DORIAN GRAY project: bridging cardiovascular disease and cognitive decline

Riccardo Proietti ^{1,*}, Marianna Adamo ², and Mauro Massucci ³

¹Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Chest and Heart Hospital, William Henry Duncan Building 6 West Derby Street, Liverpool L7 8TX, UK; ²Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, Brescia, Italy; and ³Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, Brescia, Italy

The rationale behind DORIAN GRAY

Mild cognitive impairment (MCI) and cardiovascular disease (CVD) often co-occur in older adults, with cognitive decline affecting up to one-third of patients with CVD. According to the World Health Organization, over 55 million people worldwide were living with dementia in 2021, a number projected to rise to 78 million by 2030. In Europe, the median age-standardized prevalence of CVD is ~7077 per 100 000 in men and 6026 per 100 000 in women.¹ These striking figures underscore the growing public health challenge posed by the intersection of CVD and cognitive decline. The association between MCI and CVD goes beyond shared risk factors, as CVD itself can directly contribute to MCI. Mechanistic links include cerebral infarcts affecting cognitive-critical brain regions, arterial stiffness leading to end-organ damage, and chronic cerebral hypoperfusion from low cardiac output (e.g. heart failure or atrial fibrillation) causing neuroinflammation and white matter injury.² Evidence from the European COGNITION.MATTERS HF study highlighted specific deficits in attention and verbal memory among heart failure patients, with structural correlates such as medial temporal lobe atrophy.³

Despite these findings, MCI remains underdiagnosed in CVD populations, with limited evidence-based strategies for its prevention and management.⁴ Addressing this gap requires integrating cardiovascular and cognitive data to develop personalized, proactive approaches for mitigating the dual burden of these conditions in ageing populations. In this context, the European Union, through the Horizon research and innovation action, has funded the 'Devising a Personalized Risk Stratification and Holistic Management for Prevention of Cognitive Impairment in Patients with Different Cardiovascular Phenotypes' (DORIAN GRAY) project (grant agreement no. 101156266). The consortium, led by the University of Brescia, brings together the diverse expertise of 24 leading institutions, including 12 academic partners, 5 hospitals, 5 companies, 1 scientific society, and 1 charity.

DORIAN GRAY's ambition and timeliness

The DORIAN GRAY project started in January 2025, and its programme of work is based on 17 work packages (WPs) to be delivered over a period of 5 years (Figure 1).

The DORIAN GRAY project aspires to uncover the mechanisms linking MCI with CVD and develop an integrated, personalized approach to mitigate dementia risk, promote resilience, and improve health in the ageing population. By focusing on data from CVD patients—where pathways to MCI are amplified—the project aims to identify mechanisms driving cognitive impairment and its progression in populations with cardiovascular risk.

Timely and innovative, DORIAN GRAY leverages multi-modal AI technology to harmonize clinical and real-world data from diverse sources, including wearable devices and mobile apps (Figure 1). This approach enables precise risk stratification and tailored interventions across primary, secondary, and tertiary prevention strategies. At the core of the project is the PORTRAIT platform, a cutting-edge framework that integrates genomics, biomarkers, imaging, and real-world data (RWD) to detect early signs of MCI. Complementing this is the MIRROR app, which delivers a personalized, adaptive intervention tailored to each patient. Through an avatar-based coaching system, the app dynamically adjusts cognitive and physical exercises based on user feedback and physiological data collected from smartwatches. This closed-loop approach not only enhances engagement and effectiveness but also enables real-time tracking of performance, generating valuable RWD that may serve as early markers for MCI diagnosis. In addition, the digital coaching combined with the avatar-based exergame enables the transfer of positive behaviours from the virtual world of the game to real life, avoiding the standardized, top-down format common in conventional health apps.

* Corresponding author. Email: riccardo.proietti@liverpool.ac.uk

© The Author(s) 2025. Published by Oxford University Press on behalf of the European Society of Cardiology. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

4.2 Scientific Rationale: The Heart–Brain Axis and Disease Phenomenology

Connecting the Dots: A Narrative Review of the Relationship between Heart Failure and Cognitive Impairment.

Mauro Massucci², Maria Giulia Bellicini¹, Marianna Adamo¹, Andrea Pilotto³⁻⁵, Marco Metra¹,
Alessandro Padovani³⁻⁶, Riccardo Proietti⁷

ABSTRACT

IMPORTANCE Large clinical data underscore that heart failure is independently associated to an increased risk of negative cognitive outcome and dementia. Emerging evidence suggests that cerebral hypoperfusion, stemming from reduced cardiac output and vascular pathology, may contribute to the largely overlapping vascular dementia and Alzheimer's disease. Despite these insights, cognitive outcomes remain largely overlooked in heart failure management. This narrative review outlines the prevalence and risk of cognitive impairment in heart failure patients, exploring potential shared pathophysiological mechanisms and examining the impact of heart failure therapy on cognitive deficits. Additionally, it discusses clinical implications and suggests future treatment approaches targeting therapeutic outcomes.

OBSERVATIONS Cognitive impairment is prevalent among individuals with heart failure, with reported rates varying widely depending on assessment methods. Shared pathological pathways and risk factors, including atrial fibrillation, hypertension, obesity, and type 2 diabetes mellitus, suggest a causal link. Mechanisms such as poor perfusion, microembolic events, ischemic syndromes, and cerebral inflammation contribute to this relationship. Moreover, heart failure itself may exacerbate cognitive dysfunction. This emerging understanding posits that vascular dementia and Alzheimer's disease may represent a pathophysiological continuum, driven by both the accumulation of misfolded proteins and cerebrovascular pathology due to cardiovascular dysfunction. Understanding these links is crucial for developing effective treatment strategies.

CONCLUSIONS AND RELEVANCE The complex interplay between heart failure and cognitive impairment underscores the necessity for a holistic patient care approach. Both conditions share analogous disease processes, influencing self-management and independence in patients. Prioritizing brain health in heart failure management is essential to enhance patient prognosis and general well-being.

Introduction

Heart failure (HF) is a prevalent condition, affecting at least 26 million people worldwide and posing a significant social and economic impact on public health. Based on a United States. investigation, healthcare costs for HF vary from \$441 to \$1585 per-patient-per-month (PPPM), depending on the patient's stability or a worsening HF event[1]. This burden is expected to increase as the global population ages, with HF prevalence predicted to double within 40 years[2]. Similarly, mild cognitive impairment (CI), defined as a statistical construct denoting performance on cognitive tests consistently below age- and education-specific norms [3], and its severe form, dementia, are also associated with increasing social and economic burdens. Current estimates suggest that over 55 million people worldwide are affected by CI, though many mild cases may go unrecognized [4]. In a German Health Economics study, the net annual costs of dementia vary by disease stage, ranging from €15,474 up to €41,808 [5]. Similar to HF, the prevalence of CI is higher in Western countries, primarily due to the older average age of the population. However, it is important to recognize that the lower prevalence of CI in middle and low-income countries may also be attributed to underdiagnosis, in addition to differences in population age. With the population aging in these regions as well, the number of individuals affected by dementia is projected to nearly double every 20 years worldwide, with a significant increase expected in developing countries [6]. CI and HF often co-occur in the ageing population and they probably interact each other, even if the precise pathophysiological processes underlying this association are not yet fully understood. There is recognition that HF may be considered an independent predictor, though the universality and unequivocal nature of this prediction remain subjects of ongoing research [7].

This is a non-casual association: a significant portion of dementia cases can be attributed to cardiovascular risk factors, which could potentially be prevented through cardiovascular risk modification. Furthermore, HF can lead to varying degrees of CI, influenced by genetic factors (*e.g.*, *APOE*) and environmental elements. Hence, given the presence of several of the nine 'aspects of non-casual association' discussed by Hill in 1965 (strength of association, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy), we can hypothesize a multifaceted relationship between HF and CI [8].

In this narrative review, we will firstly describe the prevalence and analyse the risk of CI, in patients with HF. Second, we will discuss the possible pathophysiological mechanisms by which HF constitutes a risk factor for CI. Third, we will summarize the literature evidence on the effects of HF therapy on cognitive dysfunction. Finally, we discuss clinical implications and future treatment approaches in terms of therapeutic targets.

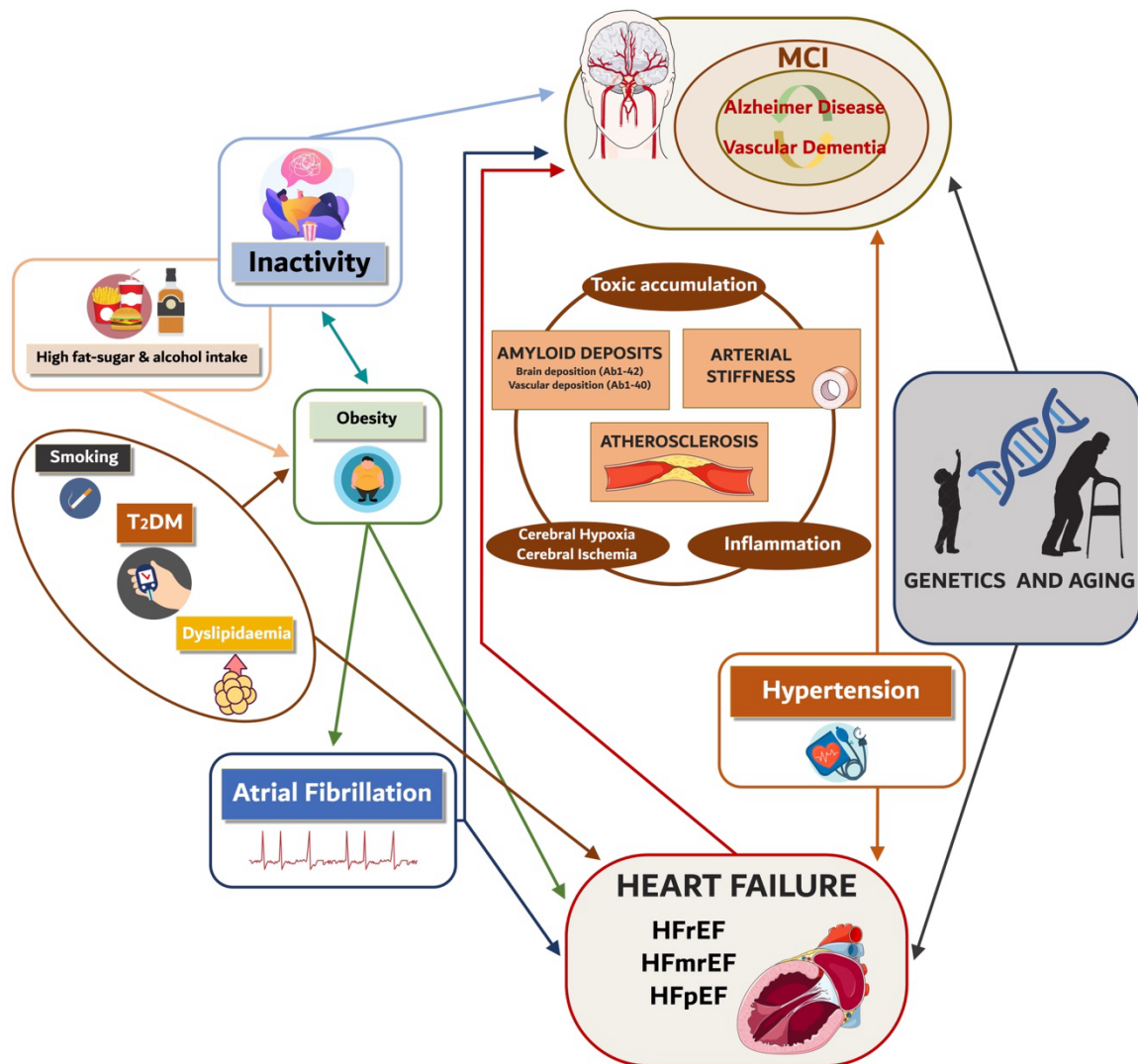
Prevalence of cognitive impairment in heart failure

According to the existing literature, CI is highly prevalent among individuals with HF, with reported rates spanning from 3% to 80%, depending on the assessment methods for cognition (**table 1**). Studies drawing from extensive cohorts, including data from International Classification of Diseases (ICD) records, consistently report a prevalence of almost 15% with some single study reporting different values. For instance, Lafo et al. observed a 36% prevalence of CI among a highly comorbid population of veterans hospitalized primarily for HF; this population had also a notably high rate of hospital readmission (30-day readmission rate: 15%, 1-year readmission rate: 59%) and mortality (30-day mortality 6%, 1-year mortality 41%)[9]. Biagi et al. reported a 35% prevalence in another extensively comorbid cohort of 1444 patients admitted with a diagnosis of chronic HF to Internal Medicine Department [10]. Conversely, three studies reported a lower dementia prevalence (3.2%, 6.9%, and 6.9%, respectively) considering a relatively low comorbid population [11–13]. Unfortunately, for the majority of these cohort studies the clinical subtypes of dementia/CI (vascular dementia vs Alzheimer's disease (AD)) are not available. The only published prospective data on mild CI (MCI) and HF originates with fully data available have been presented within the COGNITION.MATTERS HF study, following 148 patients for up to 6 years, collecting data on cognitive function, brain imaging, and inflammatory biomarkers [14]. The study's principal findings revealed that HF patients exhibited selective deficits in attention (41%) and verbal memory domains (28%), with medial temporal lobe atrophy (potentially indicating underlying AD-related pathology) identified as a probable structural correlate of cognitive dysfunction. Studies investigating the potential risk of CI development in HF patients remain limited and yield conflicting results (**table 2**). Most studies suggest a significantly higher risk compared to the general population (with HR ranging from 1.3 to 1.7), while a few show non-significant outcomes, possibly due to the relatively low number of HF patients in those studies [15–17]. In contrast, a Swedish nationwide analysis reported a potentially protective role of HF against dementia risk in patients with atrial fibrillation (AF) [18], hypothesizing that effective drug treatment in HF patients and higher mortality rates in the AF and HF subgroup might explain this unexpected result. A comprehensive meta-analysis by J.A. Cannon et al. involving 37 studies, demonstrated a significantly elevated risk of CI in HF patients (prevalence 43%) compared to matched controls without HF [19]. It's worth noting that patients affected by CI are often excluded from clinical studies, and that may underestimate the actual prevalence. Included studies, however, exhibit significant heterogeneity, with dementia prevalence ranging from 10% to 79%. Another meta-analysis by N. Li, which encompassed 119 studies, reported a pooled proportion of 41% in the 95 studies that examined CI using standardized tests. Several studies have explored the timing of cognitive decline in relation to the onset of HF. Most of these studies suggest that cognitive

decline tends to manifest after the onset of HF rather than before it. For example, Sterling et al. found that the prevalence of CI 1-18 months before HF onset was comparable to that in control subjects (14.9% [11.7-18.6%] vs. 13.4% [11.6%-15.4%], $p<0.43$) [20]. Similarly, Sun et al. reported a low prevalence of dementia in patients with newly diagnosed HF (4.2% in women and 2.5% in men) [11]. Regarding the impact of HF duration on cognitive decline, Hammond et al. reported a faster decline in Modified Mini-Mental State Examination (MMSE) scores over five years following HF onset [21]. Bressler et al. also found that the greatest six-year decline in cognitive test scores was significantly associated with an increased risk of developing HF [22]. In the ARIC study, Witt et al. observed that participants with HF had a higher prevalence of dementia (RRR = 1.60 [95% CI 1.13, 2.25]) and MCI (RRR = 1.36 [95% CI 1.12, 1.64]) at visit 5 (2011–2013), with a decline in cognition between visit 4 (1996–1998) and 5 that was greater in those who developed HF after visit 4 [23]. Long-term data from Adelborg et al. indicated an increased risk of dementia in patients with HF over 1-35 years of follow-up, with risk ratios of 1.21 (95% CI 1.18–1.24) during the first 10 years, 1.19 (95% CI 1.11–1.28) during 11-20 years, and 1.38 (95% CI 1.07–1.79) during 21-35 years [24]. Ren et al. further reported that 11.0% of patients developed new-onset dementia after a median follow-up of 4.1 years (IQR: 1.2–10.2) post-HF diagnosis, with a higher incidence in women (64%) [25]. Regarding the correlation between the severity of HF and the prevalence of CI, several studies have demonstrated a significant relationship. Lee et al. found that patients with New York Heart Association (NYHA) functional class II or higher were independently associated with cognitive decline [26]. Similarly, Brunén et al. confirmed that patients with NYHA class III-IV had a higher prevalence of CI [27]. The WARCEF trial highlighted an independent association between left ventricular ejection fraction (LVEF) and MMSE decline [28]. On the other hand, the ARIC cognition study found no significant difference in the incidence of CI between HF with reduced ejection fraction (HFrEF) and HF with preserved ejection fraction (HFpEF) patients [23]. Additionally, in acute decompensated HF, no significant disparity was observed in mean Montreal Cognitive Assessment (MoCA) scores or in the proportion of patients with MoCA scores below 26 between HFpEF and HFrEF groups [29]. The link between LVEF and cognitive function may be nonlinear, suggesting a potentially exponential association, with a stronger impact observed at lower LVEF levels compared to higher ones. Lastly, the potential misdiagnosis of depressive symptoms as cognitive impairment in older HF patients must be considered, as this could affect the accuracy of CI prevalence estimates. Some studies, have noted the difficulty in differentiating between CI and depression in these patients, highlighting the need for careful clinical evaluation [26].

Cardiovascular risk factors and cognitive impairment

Evaluation of the complex interaction between the heart and brain is necessary to address patient prognosis and well-being. The shared pathological pathways and risk factors, including AF, hypertension, obesity, and type 2 diabetes mellitus (T2DM), provide vital clues to understanding the causal link between HF and CI[30,31]. Poor perfusion, microembolic events, ischemic syndromes, cerebral inflammation, endothelial dysfunction with blood-brain barrier damage, and the presence of



amyloid deposits may collectively contribute to unravelling the intricate relationship between HF and CI (Figure 1).

Atrial fibrillation and cognitive impairment burden. The age-adjusted incidence rate of atrial fibrillation (AF), reported as 1.33 per 1000 person-years, exhibits a range from 0.13 per 1000 person-years in individuals aged under 55 years, to 7.65 per 1000 person-years in those aged 85 years or older [32]. Multiple observational studies, along with several comprehensive meta-analyses, have

Figure 15. Schematic diagram mapping the shared risk factors and underlying physiological mechanisms between heart failure and cognitive dysfunction

consistently highlighted that AF is closely linked to an elevated risk of CI and dementia. Notably, the extensive biracial population-based ARIC-NCS study, spanning two decades, found that individuals diagnosed with AF exhibited a substantially greater cognitive decline when compared to their counterparts who did not develop AF (HR 1.23, 95% Confidence Interval (CI) 1.04-1.45) [33]. Notably, even after rigorous adjustment for common confounding factors such as age, sex, education, apolipoprotein E, smoking, body mass index (BMI), arterial hypertension, type 2 diabetes mellitus (T2DM), coronary heart disease and stroke, the association remains significant, especially among younger patients. This underscores the autonomous role of AF in accelerating cognitive dysfunction, prompting further inquiry into its potential causal contribution. The primary driver of AF-induced CI is cerebral infarction. However, other proposed pathways include:

- *Cerebral Hypoperfusion*: AF unsettles the heart's atrioventricular synchrony, leading to decreased cardiac output, stroke volume, and blood pressure. Research in elderly populations with persistent AF has revealed a connection to reduced total cerebral blood flow and impaired whole brain perfusion, as assessed by phase contrast Magnetic Resonance Imaging (MRI) of the brain [34].
- *Inflammation and Systemic Atherosclerotic Vascular Disease*: inflammation is believed to promote hypercoagulability and thrombus formation, potentially increasing the risk of stroke and disrupting cerebrovascular regulation, which has links to AD and vascular dementia. Inflammation may also serve as a nonspecific marker of atherosclerotic vascular disease. Recent studies have further explored the connection between inflammation, atherosclerosis, and atrial fibrillation (AF), highlighting the potential impact on cognitive decline [35].
- *Microhemorrhage*: the relationship between the burden of cerebral microbleeds, often attributed to oral anticoagulant therapy, particularly in lobar locations (overlapping with cerebral amyloid angiopathy, CAA), and cognitive function is a subject of ongoing debate [36,37].

Recent evidence suggests that rhythm-control strategies, particularly catheter ablation, are associated with a reduced risk of CI and dementia in patients with AF. A meta-analysis by Guo et al. found that rhythm-control therapy was significantly associated with a lower risk of future dementia (HR 0.74, 95% CI 0.62–0.89) compared to rate-control strategies. Specifically, AF ablation was linked to significantly lower risks of overall dementia (HR 0.62, 95% CI 0.56–0.68), AD (HR 0.78, 95% CI 0.66–0.92), and vascular dementia (HR 0.58, 95% CI 0.42–0.80) [38]. Notably, an American study of 38,176 patients demonstrated that catheter ablation was associated with a 41% lower risk of dementia compared to antiarrhythmic drugs (1.9% vs 3.3%; HR 0.59, 95% CI 0.52–0.67, $P < .0001$) [39]. Similarly, a nationwide cohort study in Korea involving 11,726 AF patients reported a reduced risk of dementia with catheter ablation compared to antiarrhythmic or rate-control drugs alone (HR 0.73, 95% CI 0.58–0.93) [40].

Hypertension and cognitive impairment burden. Hypertension affects approximately 1.4 billion adults, 31% of the adult population, worldwide [41]. Recent findings from a comprehensive meta-analysis conducted by Qin and colleagues revealed a noteworthy association between hypertension and MCI, with an overall pooled prevalence of 30%. Interestingly, MCI prevalence in hypertensive patients demonstrated regional disparities, with rates of 26% in Asia, 40% in Europe, and 17% on a global scale. After age-stratification, the prevalence of MCI was 44% (95% CI: 1–86) in hypertensive patients under 60 years old compared to 28% (95% CI: 24–32) in those aged 60 and above. It is important to note that the younger age group's data is based on only two studies, leading to a wide confidence interval [42]. Mehra et al. reported a notably high prevalence of 66% for MCI in their study on the impact of hypertension on cognitive functions. In this study, 45.7% of participants had metabolic syndrome, a rate significantly higher than in the general population (9.2%–41%). Those with metabolic syndrome exhibited poorer cognitive performance across all domains of the MoCA, even after adjusting for age, education, depression severity, and illness duration. Lower education levels, lower income, and higher age were significantly associated with lower MoCA scores [43]. Additionally, the use of MoCA, which may lack precision in individuals with lower education levels, could contribute to the high prevalence reported [44]. Hypertension might play a pivotal role in the pathophysiology of CI, with a multifaceted impact on vascular structure. The potential connection between hypertension, HF and CI [45] is driven by a complex interplay of mechanical, cellular, and molecular factors that trigger vascular smooth muscle cell remodelling [46]. Hypertension fosters the development and accumulation of atherosclerotic plaques in key arteries, including the carotid, vertebral, and intracranial cerebral arteries. It is closely associated with or often preceded by arterial stiffening, attributed to various factors such as collagen deposition and elastin fragmentation. This stiffening elevates pulse pressure and enhances mechanical stress transmission through the cerebrovascular system, leading first to small vessel adaptive changes aimed at protecting the downstream microcirculation, and then to small penetrating arteries fibrotic thickening that are a common feature in both HF and dementia. Furthermore, microvascular rarefaction, which involves a reduction in vascular density encompassing both capillaries and arterioles, is observed in both human subjects and animal models of hypertension [47,48]. It is believed to result from the increased pressure transmitted to the microvascular bed. Given the limited presence of vessels in the white matter, this phenomenon may contribute to the development of white matter lesions. Cerebral microhaemorrhages are closely associated with hypertension [49,50] and are linked to compromised cognitive function. The progression of white matter hyperintensities and small vessel disease appears to correlate in several longitudinal studies with the duration of hypertension and the poor effectiveness

of blood pressure control [51]. Another characteristic of small vessel disease involves the enlargement of the perivascular space surrounding intracerebral arteries and veins.

Obesity, type 2 diabetes mellitus and cognitive impairment burden. In a nationwide survey conducted by the National Health and Nutrition Examination, the age-adjusted prevalence of obesity in the U.S. was recorded at 42.4% in 2017–2018, demonstrating variations across age groups: 40.0% for individuals aged 20–39, 44.8% for those aged 40–59, and 42.8% for adults aged 60 and above [52]. Additionally, a robust, dose-dependent relationship has been established, connecting higher BMI levels with an increased risk of heart failure, particularly in patients with HFpEF [53,54]. A multitude of deleterious pathological features, such as insulin resistance, gut dysbiosis, oxidative stress, inflammasome activation, and systemic inflammation, are associated with obesity and T2DM [55–57]. Each of these pathological features may contribute to neuroinflammation and brain injury. Chronic systemic inflammation is a hallmark of obesity and can be instigated by adipose tissue expansion (adipocyte hypertrophy and proliferation). Adipose tissue expansion promotes a hypoxic environment where adipocytes undergo apoptosis, causing further inflammation[58]. Adipokines are of particular interest due to their ability to modulate insulin resistance, dysregulate the gut-brain axis, and increase systemic inflammation, which may contribute to the development of neuroinflammation and dementia pathology[59]. Impaired insulin signalling may be one of the early drivers of amyloid deposition in AD, showing how this morbidity can link the pathology of T2DM and AD. Research has shown that insulin degrading enzyme can break down A β , indicating that insulin resistance could potentially contribute to changes in A β metabolism and increased amyloid pathology in AD[60]. Insulin resistance in murine models has been shown to promote amyloid precursor protein (APP) phosphorylation and to increase the formation of amyloid plaques in the brain[61], moreover insulin resistance has been shown to contribute to neuroinflammation and neurodegeneration [62,63].

Physical inactivity and cognitive impairment burden. It's crucial to underscore the pivotal role of physical inactivity in patients with HF, particularly in the context of cognitive function. Studies have compellingly demonstrated that physical inactivity is significantly associated with CI in HF patients, in terms of executive function, attention, processing speed, and cognition screening scores [64]. This evidence emphasizes the critical need for interventions aimed at reducing sedentary behaviour and increasing physical activity levels. Engaging in regular physical activity not only improves cognitive function but also holds potential for reducing depression and enhancing overall well-being [65]. Higher levels of physical activity, measured in terms of step count and time spent in moderate-vigorous activity, have consistently shown positive correlations with improved cognitive function, while lower physical activity levels have been associated with cognitive dysfunction [66].

Hyperlipidaemia and cognitive impairment burden Atherosclerosis, the hallmark of hyperlipidemia, is a systemic process that affects large and small blood vessels throughout the body, including the brain. Furthermore, the brain is highly dependent on cholesterol for membrane structure and function. Dysregulation of cholesterol metabolism has been linked to CI, and it has been suggested that cholesterol-lowering therapies could reduce the risk of cognitive dysfunction. [67,68].

Translational implications

The connection between HF and CI extends beyond shared risk factors; HF itself can potentially contribute to cognitive dysfunction. Complex interactions occur at multiple levels between AD hallmarks, such as extracellular senile plaques rich in A β peptide, intraneuronal neurofibrillary tangles (NFTs) composed of hyperphosphorylated microtubule-binding protein tau (p-tau), and key CV disease features, including neuroinflammation, cerebrovascular dysfunction, blood-brain barrier (BBB) injury, and cerebral amyloid angiopathy (CAA) [3,69]. Recent evidences suggests that AD and cerebrovascular dementia are part of a disease continuum, where intersecting pathways can favour either vascular or parenchymal amyloid deposition. The primary amyloid peptide in parenchymal lesions of AD is A β 1-42, while A β 1-40 is more prevalent in peripheral atherosclerotic lesions. Factors altering the A β 1-40/-42 ratio, like apolipoprotein E4, favour amyloid deposits in the form of cerebral amyloid angiopathy rather than parenchymal plaques. This preference for A β species in different tissues may result from substantial A β 1-40 production by platelets, plaque-invading macrophages, endothelial cells, and vascular smooth muscle cells. Additionally, *APOE* isoforms have varying effects on A β production, aggregation, and clearance [70]. Specifically, the APOEe4 allelic variant represents the most potent genetic risk factor for sporadic AD and identifies a distinct clinicopathological entity, while APOEe2 is associated with a lower AD risk [70]. Nevertheless, evidence suggests that APOEe4 is linked to compromised cerebrovascular integrity and function, contributing to blood-brain barrier dysfunction and serving as a risk factor for CI due to cerebrovascular dementia, both in the presence and absence of AD pathology.

Heart failure pharmacotherapy and cognitive impairment. The effect of pharmacological therapy for HF on cognitive dysfunction has been a topic of investigation, with specific focus on certain medications. While *anticoagulation*'s established role in reducing incident dementia in AF patients is well-known, its impact on HF patients without known AF remains uncertain. Experimental studies suggest that anticoagulant agents, such as heparin and enoxaparin, may inhibit amyloid beta neurotoxic effects through their glycosaminoglycan structure, affecting APP function and BACE1

activity [71]. The evidence regarding their effect on cognitive test performance and covert infarcts in stable CAD or peripheral artery disease patients treated with rivaroxaban and aspirin is inconclusive.

Angiotensin receptor-neprilysin inhibitors (ARNIs) are theorized to potentially impact cognitive function by affecting amyloid beta peptides in the central nervous system. Neprilysin inhibition could reduce their breakdown, while increased bradykinin levels may damage the blood-brain barrier and contribute to amyloid plaque deposition [72]. The real-world analysis of the PARADIGM-HF trial did not demonstrate these effects[73]. Similarly, inhibition of angiotensin-converting enzyme (ACE) may influence the availability of amyloid beta peptides. **ACE inhibitors (ACEIs)** have been found to increase levels of Ab1-42, while the results for Ab1-40 levels have shown inconsistency, with either an increase or no change observed. While pathophysiological hypotheses and in-vitro studies suggest potential mechanisms, clinical data also managed to shed light on this matter. For instance, in a study involving 1220 patients admitted with HF, abbreviated mental test scores improved from admission to discharge in 30% of patients after the initiation of ACEIs, compared to 22% of HF patients not receiving these drugs (odds ratio, 1.6 [95% CI, 1.2–2.1]) [74]. The role of **sodium-glucose cotransporter-2 inhibitors (SGLT2Is)** is also being explored; in a prospective study of 162 frail patients with diabetes, HFpEF, and baseline MoCA scores <26, empagliflozin monotherapy correlated with improved MoCA scores 1 month after admission, whereas treatment with insulin or metformin did not. It's important to note that the authors did not control for treatment duration or serum glucose levels [75]. These findings suggest a potential cognitive benefit associated with SGLT2Is in HF patients, and further research is needed to elucidate the mechanisms behind these effects and to explore the long-term cognitive implications of this therapy. Recent studies have shown promising outcomes for **glucagon-like peptide-1 receptor agonists (GLP-1 RAs)** in obese patients with HFpEF [76]. Four meta-analyses have been conducted to assess the influence of GLP-1 RAs on cognitive function in individuals with T2DM [77–80]. However, it's important to note that none of these analyses specifically investigated the cognitive effects in patients with HF, highlighting the need for further research to better understand the potential effects of GLP-1 RAs on cognitive function [81]. Despite the significance of CI potentially linked with HF, major pharmacological trials, including those involving ARNIs and SGLT2 inhibitors, have not extensively reported on this aspect. It may not be feasible to include dementia patients in these trials, but the inclusion of patients with MCI could provide valuable insights, considering the potential impact on adherence and treatment efficacy.

Clinical Implications of MCI in patients affected by HF:

CI and brain health in HF patients have profound clinical implications, evident in a meta-analysis of over 10,000 individuals, highlighting reduced treatment adherence, compromised self-care abilities, and a decreased likelihood of seeking assistance among those with CI [82]. While HF guidelines underscore the importance of recognizing cognitive impairment, especially in the context of frailty, communication challenges, and end-of-life decisions, they currently lack specific recommendations for routine screening or diagnosis in clinical practice. Additionally, insights from the Registro de Insuficiencia Cardíaca (RICA) suggest that HF patients with severe CI face heightened mortality and morbidity risks, characterized by advanced age, increased comorbidity burden, lower survival rates, and a higher incidence of death or readmission at 1 year [27]. Nevertheless, acknowledging this association is crucial as it influences the risk/benefit ratio of various therapeutic interventions. The efficacy of pharmacotherapy and procedural interventions may be constrained if patients have limited life expectancy to derive benefits. For instance, if life expectancy is less than one year, the probability of benefit from interventions such as implantable cardioverter-defibrillators or transcatheter mitral/aortic valve procedures is low, and hence not warranted [83].

Beyond clinical implications, the economic impact of cognitive impairment in HF warrants attention, emphasizing the need for a comprehensive bioeconomical assessment to address the multifaceted challenges posed by this condition.

To address a promptly diagnose of MCI in patients with HF, van Nieuwkerk and colleagues have proposed a two-step algorithm [84]. The first step involves inquiring about substantial cognitive decline noted by patients or their relatives over the past year, followed by assessing its impact on daily life. A positive response triggers cognitive screening, as some patients with severe deficits may not perceive their cognitive problems. Clinical suspicion, raised by unexplained falls, medication errors, history of delirium, or depressive symptoms, also warrants screening, especially in patients over 65. The second step involves a multidomain cognitive screening test tailored to the patient's baseline cognitive level. Given that executive functioning and memory are commonly affected in cardiac patients, screening tools should cover these domains. The MoCA and MMSE are commonly used for global cognition assessment. MMSE, with a cutoff of 24 points, exhibits good sensitivity and negative predictive value but lower specificity for dementia. MoCA, designed for MCI, is more sensitive but has lower specificity for dementia [85]. An optimal MoCA cutoff may be 26, showing excellent sensitivity and negative predictive value but poorer specificity for dementia. The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) is useful for proxies in assessing cognitive trajectories [73]. Multidomain cognitive screening tool, such as MoCA or the MMSE or similar digital cognitive testing [86] stratify risk and inform the need for referral to specialist memory

services. HF patients might be affected by multiple domains CI (including attention, executive function, language, memory, and visual-spatial capacities). The inclusion of standardized neurocognitive outcomes within cardiovascular trials could serve as a beacon for identifying high-risk patients. NeuroARC, with its commitment to standardized neuropsychological endpoint definitions, advocates for the integration of cognitive screening at each trial visit, possibly leveraging the MoCA as a valuable tool [87]. The current landscape lacks reliable imaging and blood biomarkers for identifying individuals at risk for AD or cognitive decline[88,89]. Integrating such biomarkers into screening strategies could significantly advance the identification of high-risk patients and improve overall cognitive assessment in the HF population [75].

Management

A recent statement from the American Heart Failure Society highlights that there are no specific interventions proven to improve cognition or delay the progression of CI in patients with HF [90]. Potential approaches include the treatment of contributing factors, the promotion of physical activity, leaving specific advanced neurologic pharmacotherapies for selected cases[91] (**figure 2**).

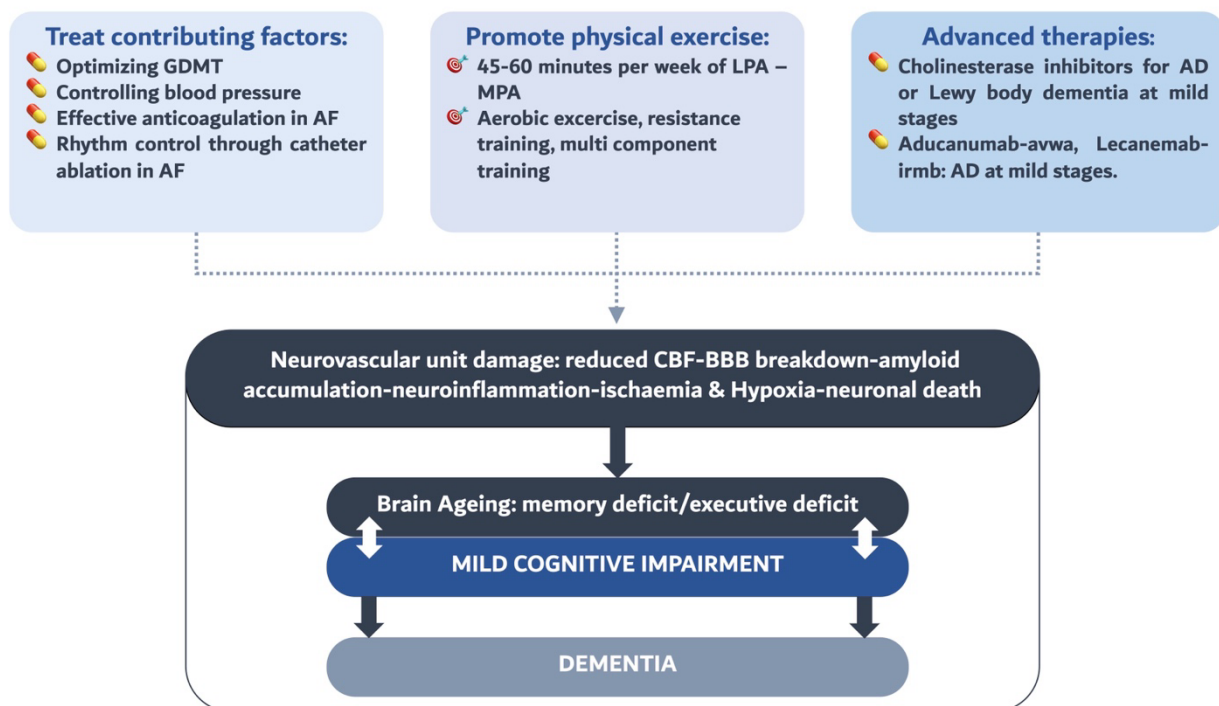


Figure 16. Potential Approaches for Managing Cognitive Dysfunction in Heart Failure Patients. This diagram outlines various pharmacological and non-pharmacological strategies for addressing cognitive dysfunction in individuals with heart failure.

- 1) Given the intricate nature of medication schedules and the prevalence of medication-related issues among individuals with HF and CI, physicians must meticulously assess and reconcile medication plans. Implementing strategies to enhance outcomes in this context may involve deprescribing, a

supervised process of medication discontinuation. Candidate medications for deprescribing might include those known to exacerbate HF and/or listed in criteria such as the Beers criteria [92], which identifies medications with potential risks outweighing benefits in older adults, particularly concerning cognition. Additionally, treating comorbidities can further improve outcomes in this patient population (*i.e.* pressure control in hypertensive patients and rhythm control in AF patients [93]).

- 2) Despite the pleiotropic effect of exercise training in HF, including enhancements in exercise capacity, cardiac function, peripheral effects, and quality of life, limited research has explored the impact of physical exercise on individuals with HF and MCI. Tanne et al. demonstrated cognitive improvements in 20 severe HF patients after an 18-week supervised exercise program [94]. Redwine et al. reported similar enhancements in cognitive performance in 69 HF participants with tai chi and resistance band exercises over 16 weeks [95]. However, Kitzman et al. did not observe cognitive improvements in 349 HF patients following a 12-week rehabilitation program (multi domain physical rehabilitation program emphasising strength, balance, mobility, and endurance) [96]. Conversely, Gary et al. found improved verbal memory in 69 HF participants with combined exercise and cognitive training over three months [97]. The aforementioned studies have poorly defined interventions, and short follow-up durations, thereby posing a high risk of bias.
- 3) New pharmacotherapeutic agents have emerged for addressing distinct categories of cognitive dysfunction. Cholinesterase inhibitors (such as donepezil, rivastigmine, and galantamine) are presently advocated for managing MCI and dementia associated with AD. However, the utility of cholinesterase inhibitors in cognitive impairment unrelated to AD (for example in vascular CI) is not supported by evidence; this raises concerns regarding heightened risk of adverse effects (*e.g.*, cardiac and gastrointestinal), alongside varying levels of evidence quality. Consequently, consensus guidelines do not advocate for the use of cholinesterase inhibitors in treating cognitive impairment not associated with AD due to the off-label nature and lack of evidence-based support for such usage [98]. Recently, disease-altering monoclonal antibodies have gained approval for treating individuals with MCI and dementia arising from AD. Phase 3 trials have demonstrated the efficacy of monoclonal antibodies (aducanumab-avwa; lecanemab-irmb) in diminishing amyloid-beta plaques within the brain, resulting in moderate alleviation of CI [99,100]. However, both treatments are associated with an increased vulnerability to amyloid-related imaging abnormalities, including cerebral microhemorrhages, cerebral macrohemorrhages, superficial siderosis, brain edema, or sulcal effusion. Subjects with HF were excluded from these clinical trials, considering the widespread use of anticoagulation in HF patients. However, the benefit and targets of disease-modifying treatment beyond classical AD is still ongoing and definitively need

to be updated according to new findings and prospective results.

Conclusions

In conclusion, the intricate relationship between HF and CI underscores the urgent need for a comprehensive approach to patient care. Both conditions are subject to similar disease processes and pathophysiological mechanisms, emphasizing their close interplay. The profound impact of CI on HF self-care and independence further emphasizes the importance of addressing cognitive health in HF management, with significant implications for patient prognosis and quality of life. This accentuates the critical need for treatment strategies capable of addressing the cognitive dysfunction associated with HF, leading to improved patient brain health and well-being.

References

1. Butler Laurence Djatche Baanie Sawhney Sreya Chakladar Lingfeng Yang Joanne E Brady Mei Yang JM. Clinical and Economic Burden of Chronic Heart Failure and Reduced Ejection Fraction Following a Worsening Heart Failure Event. *Adv Ther.* n.d.;37.
2. Savarese G, Lund LH. Global Public Health Burden of Heart Failure. *Card Fail Rev.* 2017;03:7.
3. Frisoni GB, Altomare D, Ribaldi F, et al. Dementia prevention in memory clinics: recommendations from the European task force for brain health services. *Lancet Reg Health Eur.* 2023 Jan 31;26:100576. doi: 10.1016/j.lanepe.2022.100576. PMID: 36895446.
4. Feigin VL, Vos T, Alahdab F, Amit AML et al. Burden of Neurological Disorders across the US from 1990-2017: A Global Burden of Disease Study. *JAMA Neurol.* 2021;78:165–76.
5. Leicht H, Heinrich S, Heider D, Bachmann C, Bickel H, van den Bussche H, Fuchs A, Lupp M, Maier W, Mösch E, Pentzek M, Rieder-Heller SG, Tebarth F, Werle J, Weyerer S, Wiese B, Zimmermann T, König HH. Net costs of dementia by disease stage. *Acta Psychiatr Scand.* 2011;124:384–95.
6. Padovani A, Pilotto A. Looking at the burden of neurological disorders in Europe. *Lancet Public Health.* 2020;5:e523.
7. Kuipers S, Greving JP, Brunner-La Rocca HP, Gottesman RF, van Oostenbrugge RJ, Williams NL, Jan Biessels G, Jaap Kappelle L. Risk evaluation of cognitive impairment in patients with heart failure: A call for action. *IJC Heart and Vasculature.* 2022;43.
8. Fedak KM, Bernal A, Capshaw ZA, Gross S. Applying the Bradford Hill criteria in the 21st century: How data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol.* 2015;12.
9. Lafo J, Singh M, Jiang L, Correia S, Madrigal C, Clements R, Wu WC, Erqou S, Rudolph JL. Outcomes in heart failure patients discharged to skilled nursing facilities with delirium. *ESC Heart Fail.* 2022;9:1891–900.
10. Biagi P, Nardi R, Baldo CI, Scanelli G, Panuccio D, Bonanome A, Miserocchi F, Lusiani L, Bonizzoni E, Campanini M, Nozzoli C, Vescovo G. Clinical characteristics of very old patients hospitalized in internal medicine wards for heart failure: A sub-analysis of the FADOI-CONFINE Study Group. *Italian Journal of Medicine.* 2014;8:19–28.
11. Sun LY, Tu J V., Coutinho T, Turek M, Rubens FD, McDonnell L, Tulloch H, Eddeen AB, Mielniczuk LM. Sex differences in outcomes of heart failure in an ambulatory, population-based cohort from 2009 to 2013. *Can Med Assoc J.* 2018;190:E848–54.
12. Fujiki S, Kashimura T, Okura Y, Koder K, Watanabe H, Tanaka K, Bannai S, Hatano T, Tanaka T, Kitamura N, Minamino T, Inomata T. Incidence and Risk Factors of Future Need for Long-Term Care Insurance in Japanese Elderly Patients with Left Ventricular Systolic Dysfunction. *Circulation Journal.* 2022;86:158–65.
13. Carter P, Reynolds J, Carter A, Potluri S, Uppal H, Chandran S, Potluri R. The impact of psychiatric comorbidities on the length of hospital stay in patients with heart failure. *Int J Cardiol.* 2016;207:292–6.
14. Villringer A, Laufs U. Heart failure, cognition, and brain damage. *Eur Heart J.* 2021;42:1579–81.
15. Haring B, Leng X, Robinson J, Johnson KC, Jackson RD, Beyth R, Wactawski-Wende J, von Ballmoos MW yler, Goveas JS, Kuller LH, Wassertheil-Smoller S. Cardiovascular disease and cognitive decline in postmenopausal women: results from the Women's Health Initiative Memory Study. *J Am Heart Assoc.* 2013;2:e000369.
16. Peters R, Poulter R, Beckett N, Forette F, Fagard R, Potter J, Swift C, Anderson C, Fletcher A, Bulpitt CJ. Cardiovascular and biochemical risk factors for incident dementia in the hypertension in the very elderly trial. *J Hypertens.* 2009;27:2055–62.
17. de Bruijn RFAG, Bos MJ, Portegies MLP, Hofman A, Franco OH, Koudstaal PJ, Ikram MA. The potential for prevention of dementia across two decades: The prospective, population-based Rotterdam Study. *BMC Med.* 2015;13.
18. Wändell P, Carlsson AC, Holzmänn MJ, Årnlöv J, Sundquist J, Sundquist K. The association between relevant co-morbidities and prevalent as well as incident heart failure in patients with atrial fibrillation. *J Cardiol.* 2018;72:26–32.
19. Cannon JA, Moffitt P, Perez-Moreno AC, Walters MR, Broomfield NM, McMurray JJV, Quinn TJ. Cognitive Impairment and Heart Failure: Systematic Review and Meta-Analysis. *J Card Fail.* 2017;23:464–75.
20. Sterling MR, Jannat-Khah D, Bryan J, Banerjee S, McClure LA, Wadley VG, Unverzagt FW, Levitan EB, Goyal P, Peterson JC, Manly JJ, Levine DA, Safford MM. The Prevalence of Cognitive Impairment Among Adults With Incident Heart Failure: The “Reasons for Geographic and Racial Differences in Stroke” (REGARDS) Study. *J Card Fail.* 2019;25:130–6.
21. Hammond CA, Blades NJ, Chaudhry SI, Dodson JA, Longstreth WT, Heckbert SR, Psaty BM, Arnold AM, Dublin S, Sitlani CM, Gardin JM, Thielke SM, Nanna MG, Gottesman RF, Newman AB, Thacker EL. Long-Term Cognitive Decline After Newly Diagnosed Heart Failure: Longitudinal Analysis in the CHS (Cardiovascular Health Study). *Circ Heart Fail.* 2018;11:e004476.
22. Bressler J, Knopman DS, Sharrett AR, Gottesman RF, Penman A, Chang PP, Rosamond WD, Boerwinkle E, Mosley TH. Incident Heart Failure and Cognitive Decline: The Atherosclerosis Risk in Communities Study. *J Card Fail.* 2017;23:47–55.

23. Witt LS, Rotter J, Stearns SC, Gottesman RF, Kucharska-Newton AM, Richey Sharrett A, Wruck LM, Bressler J, Sueta CA, Chang PP. Heart Failure and Cognitive Impairment in the Atherosclerosis Risk in Communities (ARIC) Study. *J Gen Intern Med*. 2018;33:1721–8.
24. Adelborg K, Horváth-Puhó E, Ording A, Pedersen L, Sørensen HT, Henderson VW. Heart failure and risk of dementia: a Danish nationwide population-based cohort study. *Eur J Heart Fail*. 2017;19:253–60.
25. Ren Q wen, Katherine Teng TH, Tse YK, Tay WT, Li HL, Tromp J, Yu SY, Hung D, Wu MZ, Chen C, Yuk Yuen JK, Huang JY, Ouwerkerk W, Li XL, Teramoto K, Chandramouli C, Tse HF, Lam CSP, Yiu KH. Incidence, Clinical Correlates, and Prognostic Impact of Dementia in Heart Failure: A Population-Based Cohort Study. *JACC: Asia*. 2023;3:108–19.
26. Lee TC, Qian M, Liu Y, Graham S, Mann DL, Nakanishi K, Teerlink JR, Lip GYH, Freudenberger RS, Sacco RL, Mohr JP, Labovitz AJ, Ponikowski P, Lok DJ, Matsumoto K, Estol C, Anker SD, Pullicino PM, Buchsbaum R, Levin B, Thompson JLP, Homma S, Di Tullio MR. Cognitive Decline Over Time in Patients With Systolic Heart Failure: Insights From WARCEF. *JACC Heart Fail*. 2019;7:1042–53.
27. García Bruñén JM, Povar Echeverría M, Díez-Manglano J, Manzano L, Trullàs JC, Romero Requena JM, Salamanca Bautista M *P, González Franco Á, Cepeda Rodrigo JM, Montero-Pérez-Barquero M. Cognitive impairment in patients hospitalized for congestive heart failure: data from the RICA Registry. *Intern Emerg Med*. 2021;16:141–8.
28. Homma S, Thompson JLP, Pullicino PM, Levin B, Freudenberger RS, Teerlink JR, Ammon SE, Graham S, Sacco RL, Mann DL, Mohr JP, Massie BM, Labovitz AJ, Anker SD, Lok DJ, Ponikowski P, Estol CJ, Lip GYH, Di Tullio MR, Sanford AR, Mejia V, Gabriel AP, del Valle ML, Buchsbaum R. Warfarin and Aspirin in Patients with Heart Failure and Sinus Rhythm. *New England Journal of Medicine*. 2012;366:1859–69.
29. Warraich HJ, Kitzman DW, Whellan DJ, Duncan PW, Mentz RJ, Pastva AM, Nelson MB, Upadhyay B, Reeves GR. Physical Function, Frailty, Cognition, Depression, and Quality of Life in Hospitalized Adults ≥60 Years With Acute Decompensated Heart Failure With Preserved Versus Reduced Ejection Fraction. *Circ Heart Fail*. 2018;11:e005254.
30. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, Brayne C, Burns A, Cohen-Mansfield J, Cooper C, Costafreda SG, Dias A, Fox N, Gitlin LN, Howard R, Kales HC, Kivimäki M, Larson EB, Ogunniyi A, Orgeta V, Ritchie K, Rockwood K, Sampson EL, Samus Q, Schneider LS, Selbæk G, Teri L, Mukadam N. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*. 2020;396:413–46.
31. Kivipelto M, Helkala EL, Hänninen T, Laakso MP, Hallikainen M, Alhainen K, Soininen H, Tuomilehto J, Nissinen A. Midlife vascular risk factors and late-life mild cognitive impairment: A population-based study. *Neurology*. 2001;56:1683–9.
32. Lane DA, Skjøth F, Lip GYH, Larsen TB, Kotecha D. Temporal trends in incidence, prevalence, and mortality of atrial fibrillation in primary care. *J Am Heart Assoc*. 2017;6.
33. Chen LY, Norby FL, Gottesman RF, Mosley TH, Soliman EZ, Agarwal SK, Loefer LR, Folsom AR, Coresh J, Alonso A. Association of atrial fibrillation with cognitive decline and dementia over 20 years: The ARIC-NCS (Atherosclerosis Risk in Communities Neurocognitive Study). *J Am Heart Assoc*. 2018;7.
34. Gardarsdóttir M, Sigurdsson S, Aspelund T, Rokita H, Launer LJ, Gudnason V, Arnar DO. Atrial fibrillation is associated with decreased total cerebral blood flow and brain perfusion. *Europace*. 2018;20:1252–8.
35. Nso N, Bookani KR, Metzl M, Radparvar F. Role of inflammation in atrial fibrillation: A comprehensive review of current knowledge. *J Arrhythm*. 2021;37:1–10.
36. Conen D, Rodondi M, Müller A et al. . Relationships of Overt and Silent Brain Lesions With Cognitive Function in Patients With Atrial Fibrillation. *J Am Coll Cardiol*. 2019;73:989–99.
37. Saito T, Kawamura Y, Sato N, Kano K, Takahashi K, Asanome A, Sawada J, Katayama T, Hasebe N. Non-vitamin K antagonist oral anticoagulants do not increase cerebral microbleeds. *Journal of Stroke and Cerebrovascular Diseases*. 2015;24:1373–7.
38. Guo J, Liu Y, Jia J, Lu J, Wang D, Zhang J, Ding J, Zhao X. Effects of rhythm-control and rate-control strategies on cognitive function and dementia in atrial fibrillation: a systematic review and meta-analysis. *Age Ageing*. 2024;53.
39. Zeitler EP, Bunch TJ, Khanna R, Fan X, Iglesias M, Russo AM. Comparative risk of dementia among patients with atrial fibrillation treated with catheter ablation versus anti-arrhythmic drugs. *Am Heart J*. 2022;254:194–202.
40. Kim D, Yang PS, Joung B. Prevention of Dementia in patients with atrial fibrillation. *Korean Circ J*. 2021;51:308–19.
41. Forouzanfar MH, Liu P, Roth G et al. Global burden of hypertension and systolic blood pressure of at least 110 to 115mmHg, 1990-2015. *JAMA - Journal of the American Medical Association*. 2017;317:165–82.
42. Qin J, He Z, Wu L, Lin Q, Lin Y, Zheng L. Prevalence of mild cognitive impairment in patients with hypertension: a systematic review and meta-analysis. *Hypertens Res*. 2021;1251–60.
43. Mehra A, Suri V, Kumari S, Avasthi A, Grover S. Association of mild cognitive impairment and metabolic syndrome in patients with hypertension. *Asian J Psychiatr*. 2020;53.
44. Zhang S, Qiu Q, Qian S, Lin X, Yan F, Sun L, Xiao S, Wang J, Fang Y, Li X. Determining Appropriate Screening Tools and Cutoffs for Cognitive Impairment in the Chinese Elderly. *Front Psychiatry*. 2021;12.
45. Santisteban MM, Iadecola C, Carnevale D. Hypertension, Neurovascular Dysfunction, and Cognitive Impairment. *Hypertension*. 2023;80:22–34.
46. Webb AJS, Werring DJ. New Insights into Cerebrovascular Pathophysiology and Hypertension. *Stroke*. 2022;53:1054–64.
47. Hu X, Michael De Silva T, Chen J, Faraci FM. Cerebral Vascular Disease and Neurovascular Injury in Ischemic Stroke. *Circ Res*. 2017;120:449–71.
48. Ungvari Z, Toth P, Tarantini S, Prodan CI, Sorond F, Merkely B, Csiszar A. Hypertension-induced cognitive impairment: from pathophysiology to public health. *Nat Rev Nephrol*. 2021;17:639–54.
49. Poels MMF, Ikram MA, Van Der Lugt A, Hofman PA, Niessen WJ, Krestin GP, Breteler MMB, Vernooij MW. Cerebral microbleeds are associated with worse cognitive function The Rotterdam Scan Study. 2012.
50. Romero JR, Beiser A, Himali JJ, Shoamanesh A, DeCarli C, Seshadri S. Cerebral microbleeds and risk of incident dementia: the Framingham Heart Study. *Neurobiol Aging*. 2017;54:94–9.
51. McCaw Z, Kim DH, Wei L-J. Analysis of Long-term Benefits of Intensive Blood Pressure Control. *JAMA*. 2019.
52. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017-2018 Key findings Data from the National Health and Nutrition Examination Survey. 2017.
53. Pandey A, LaMonte M, Klein L, Ayers C, Psaty BM, Eaton CB, Allen NB, de Lemos JA, Carnethon M, Greenland P, Berry JD. Relationship Between Physical Activity, Body Mass Index, and Risk of Heart Failure. *J Am Coll Cardiol*. 2017;69:1129–42.
54. Adamson C, Kondo T, Jhund PS, De Boer RA, Cabrera Honorio JW, Claggett B, Desai AS, Alcocer Gamba MA, Al Habeeb W, Hernandez AF, Inzucchi SE, Kosiborod MN, Lam CSP, Langkilde AM, Lindholm D, Bachus E, Litwin SE, Martinez F, Petersson M, Shah SJ, Vaduganathan M, Nguyen Vinh P, Wilderäng U, Solomon SD, McMurray JJV. Dapagliflozin for heart failure according to body mass index: the DELIVER trial. *Eur Heart J*. 2022;43:4406–17.

55. Ellulu MS, Patimah I, Khaza'ai H, Rahmat A, Abed Y. Obesity & inflammation: The linking mechanism & the complications. *Archives of Medical Science*. 2017;13:851–63.
56. Loffredo L, Ettorre E, Zicari AM, Inghilleri M, Nocella C, Perri L, Spalice A, Fossati C, De Lucia MC, Pigozzi F, Cacciafesta M, Violi F, Carnevale R. Oxidative Stress and Gut-Derived Lipopolysaccharides in Neurodegenerative Disease: Role of NOX2. *Oxid Med Cell Longev*. 2020;2020.
57. YU WEI, GAO D, JIN WEN, WANG Z, LI YAN, PENG X, CONG Y, LI C, ZHAO A, LIU S, QI S. Intestinal Flora Dysbiosis Aggravates Cognitive Dysfunction Associated With Neuroinflammation in Heart Failure. *J Card Fail*. 2020;26:885–94.
58. Lindhorst A, Raulien N, Wieghofer P, Eilers J, Rossi FMV, Bechmann I, Gericke M. Adipocyte death triggers a pro-inflammatory response and induces metabolic activation of resident macrophages. *Cell Death Dis*. 2021;12.
59. Kiliaan AJ, Arnoldussen IAC, Gustafson DR. Review Adipokines: a link between obesity and dementia? vol. 13. 2014.
60. Zhao Z, Ho L, Wang J, Qin W, Festa ED, Mobbs C, Hof P, Rocher A, Masur S, Haroutunian V, Pasinetti GM. Connective tissue growth factor (CTGF) expression in the brain is a downstream effector of insulin resistance-associated promotion of Alzheimer's disease β -amyloid neuropathology. *The FASEB Journal*. 2005;19:2081–2.
61. Czczor JK, Genders AJ, Aston-Mourney K, Connor T, Hall LG, Hasebe K, Ellis M, De Jong KA, Henstridge DC, Meikle PJ, Febbraio MA, Walder K, McGee SL. APP deficiency results in resistance to obesity but impairs glucose tolerance upon high fat feeding. *Journal of Endocrinology*. 2018;237:311–22.
62. Forny-Germano L, De Felice FG, Do Nascimento Vieira MN. The role of leptin and adiponectin in obesity-associated cognitive decline and Alzheimer's disease. *Front Neurosci*. 2019;13.
63. Sweeney MD, Sagare AP, Zlokovic B V. Blood-brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. *Nat Rev Neurol*. 2018;14:133–50.
64. Alosco ML, Spitznagel MB, Sweet LH, Josephson R, Hughes J, Gunstad J. Atrial fibrillation exacerbates cognitive dysfunction and cerebral perfusion in heart failure. *PACE - Pacing and Clinical Electrophysiology*. 2015;38:178–86.
65. Mahindru A, Patil P, Agrawal V. Role of Physical Activity on Mental Health and Well-Being: A Review. *Cureus*. 2023.
66. Arnold SE, Arvanitakis Z, Macauley-Rambach SL, Koenig AM, Wang HY, Ahima RS, Craft S, Gandy S, Buettner C, Stoeckel LE, Holtzman DM, Nathan DM. Brain insulin resistance in type 2 diabetes and Alzheimer disease: Concepts and conundrums. *Nat Rev Neurol*. 2018;14:168–81.
67. Song G, Liu J, Zhao Z, Yu Y, Tian H, Yao S, Li G, Qin S. Simvastatin reduces atherogenesis and promotes the expression of hepatic genes associated with reverse cholesterol transport in apoE-knockout mice fed high-fat diet. *Lipids Health Dis*. 2011;10.
68. Getz GS, Reardon CA. Do the ApoE^{-/-} and Ldlr^{-/-} mice yield the same insight on atherogenesis? *Arterioscler Thromb Vasc Biol*. 2016;36:1734–41.
69. Kumar A, Fontana IC, Nordberg A. Reactive astrogliosis: A friend or foe in the pathogenesis of Alzheimer's disease. *J Neurochem*. 2023;164:309–24.
70. Yamazaki Y, Zhao N, Caulfield TR, Liu CC, Bu G. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nat Rev Neurol*. 2019;15:501–18.
71. Mycroft-West CJ, Devlin AJ, Cooper LC, Procter P, Miller GJ, Fernig DG, Guerrini M, Guimond SE, Lima MA, Yates EA, Skidmore MA. Inhibition of BACE1, the β -secretase implicated in Alzheimer's disease, by a chondroitin sulfate extract from *Sardina pilchardus*. *Neural Regen Res*. 2020;15:1546–53.
72. Campbell DJ. Long-term neprilysin inhibition-implications for ARNIs. *Nat Rev Cardiol*. 2017;14:171–86.
73. McMurray JJV, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR, Rouleau JL, Shi VC, Solomon SD, Swedberg K, Zile MR. Angiotensin–Neprilysin Inhibition versus Enalapril in Heart Failure. *New England Journal of Medicine*. 2014;371:993–1004.
74. Zuccalà G, Onder G, Marzetti E, Lo Monaco MR, Cesari M, Cocchi A, Carboni P, Bernabei R. Use of angiotensin-converting enzyme inhibitors and variations in cognitive performance among patients with heart failure. *Eur Heart J*. 2005;26:226–33.
75. Mone P, Lombardi A, Gambardella J, Pansini A, Macina G, Morgante M, Frullone S, Santulli G. Empagliflozin Improves Cognitive Impairment in Frail Older Adults With Type 2 Diabetes and Heart Failure With Preserved Ejection Fraction. *Diabetes Care*. 2022;45:1247–51.
76. Kosiborod MN, Abildstrøm SZ, Borlaug BA, Butler J, Christensen L, Davies M, Hovingh KG, Kitzman DW, Lindegaard ML, Møller DV, Shah SJ, Treppendahl MB, Verma S, Petrie MC. Design and Baseline Characteristics of STEP-HFpEF Program Evaluating Semaglutide in Patients With Obesity HFpEF Phenotype. *JACC Heart Fail*. 2023;11:1000–10.
77. Bi Z, Wang L, Wang W. Evaluating the effects of glucagon-like peptide-1 receptor agonists on cognitive function in Alzheimer's disease: A systematic review and meta-analysis. *Advances in Clinical and Experimental Medicine*. 2023;32.
78. Tian S, Jiang J, Wang J, Zhang Z, Miao Y, Ji X, Bi Y. Comparison on cognitive outcomes of antidiabetic agents for type 2 diabetes: A systematic review and network meta-analysis. *Diabetes Metab Res Rev*. 2023;39.
79. Luan S, Cheng W, Wang C, Gong J, Zhou J. Impact of glucagon-like peptide 1 analogs on cognitive function among patients with type 2 diabetes mellitus: A systematic review and meta-analysis. *Front Endocrinol (Lausanne)*. 2022;13.
80. Jin Y, Zhao H, Hou Y, Song G. The effects of dipeptidyl peptidase-4 inhibitors and glucagon-like peptide 1 receptor agonists on cognitive functions in adults with type 2 diabetes mellitus: a systematic review and meta-analysis. *Acta Diabetol*. 2020;57:1129–44.
81. Khan MS, Fonarow GC, McGuire DK, Hernandez AF, Vaduganathan M, Rosenstock J, Handelsman Y, Verma S, Anker SD, McMurray JJV, Kosiborod MN, Butler J. Glucagon-Like Peptide 1 Receptor Agonists and Heart Failure: The Need for Further Evidence Generation and Practice Guidelines Optimization. *Circulation*. 2020;142:1205–18.
82. Lovell J, Pham T, Noaman SQ, Davis MC, Johnson M, Ibrahim JE. Self-management of heart failure in dementia and cognitive impairment: A systematic review. *BMC Cardiovasc Disord*. 2019;19.
83. Massucci M, Adamo M, Rosato S, Seccareccia F, Barbanti M, Biancari F, Tarantini G, Immè S, Vignali L, Badoni G, Ussia GP, Costa G, Baiocchi M, Metra M, Baglio G, D'Errigo P. Functional and metabolic frailty predicts mortality in patients undergoing TAVI: Insights from the OBSERVANT II study. *Eur J Intern Med*. 2022;106:90–6.
84. van Nieuwkerk AC, Delewi FJ, Mullers FJ, Daemen M, Biessels GJ. Cognitive Impairment in Patients With Cardiac Disease: Implications for Clinical Practice. *Stroke*. 2023;54:2181–91.
85. Tsoi KKF, Chan JYC, Hirai HW, Wong SYS, Kwok TCY. Cognitive tests to detect dementia a systematic review and meta-analysis. *JAMA Intern Med*. 2015;175:1450–8.
86. Vyshedskiy A, Netson R, Fridberg E, Jagadeesan P, Arnold M, Barnett S, Gondalia A, Maslova V, de Torres L, Ostrovsky S, Durakovic D, Savchenko A, McNett S, Kogan M, Piryatinsky I, Gold D. Boston cognitive assessment (BOCA) — a comprehensive self-administered smartphone- and computer-based at-home test for longitudinal tracking of cognitive performance. *BMC Neurol*. 2022;22:92.
87. Lansky AJ, Messé SR, Brickman AM et al. Proposed Standardized Neurological Endpoints for Cardiovascular Clinical Trials: An Academic Research Consortium Initiative. *J Am Coll Cardiol*. 2017;69:679–91.
88. Hansson O, Blennow K, Zetterberg H, Dage J. Blood biomarkers for Alzheimer's disease in clinical practice and trials. *Nat Aging*. 2023;3:506–19.

89. Ashton NJ, Janelidze S, Mattsson-Carlgren N, et al. Differential roles of A β 42/40, p-tau231 and p-tau217 for Alzheimer's trial selection and disease monitoring. *Nat Med*. 2022;28:2555–62.
90. GOYAL P, DIDOMENICO RJ, PRESSLER SJ, IBEH C, WHITE-WILLIAMS C, ALLEN LA, GORODESKI EZ, Albert N, Fudim M, Lekavich C, Watson K, Gulati S, Kalogeropoulos A, Lewsey S. Cognitive Impairment in Heart Failure: A Heart Failure Society of America Scientific Statement. *J Card Fail*. 2024;30:488–504.
91. Rosenberg A, Mangialasche F, Ngandu T, Solomon A, Kivipelto M. Multidomain Interventions to Prevent Cognitive Impairment, Alzheimer's Disease, and Dementia: From FINGER to World-Wide FINGERS. *J Prev Alzheimers Dis*. 2020;7:29–36.
92. Jaber D, Vargas F, Nguyen L, Ringel J, Zarzuela K, Musse M, Kwak MJ, Levitan EB, Maurer MS, Lachs MS, Safford MM, Goyal P. Prescriptions for Potentially Inappropriate Medications from the Beers Criteria Among Older Adults Hospitalized for Heart Failure. *J Card Fail*. 2022;28:906–15.
93. Bunch TJ, Crandall BG, Weiss JP, May HT, Bair TL, Osborn JS, Anderson JL, Muhlestein JB, Horne BD, Lappe DL, Day JD. Patients treated with catheter ablation for atrial fibrillation have long-term rates of death, stroke, and dementia similar to patients without atrial fibrillation. *J Cardiovasc Electrophysiol*. 2011;22:839–45.
94. Tanne D, Freimark D, Poreh A, Merzeliak O, Bruck B, Schwammenthal Y, Schwammenthal E, Motro M, Adler Y. Cognitive functions in severe congestive heart failure before and after an exercise training program. *Int J Cardiol*. 2005;103:145–9.
95. Redwine LS, Pung MA, Wilson K, Bangen KJ, Delano-Wood L, Hurwitz B. An exploratory randomized sub-study of light-to-moderate intensity exercise on cognitive function, depression symptoms and inflammation in older adults with heart failure. *J Psychosom Res*. 2020;128.
96. Kitzman DW, Whellan DJ, Duncan P, Pastva AM, Mentz RJ, Reeves GR, Nelson MB, Chen H, Upadhyay B, Reed SD, Espeland MA, Hewston L, O'Connor CM. Physical Rehabilitation for Older Patients Hospitalized for Heart Failure. *New England Journal of Medicine*. 2021;385:203–16.
97. Gary RA, Paul S, Corwin E, Butts B, Miller AH, Hepburn K, Williams B, Waldrop-Valverde D. Exercise and Cognitive Training as a Strategy to Improve Neurocognitive Outcomes in Heart Failure: A Pilot Study. *American Journal of Geriatric Psychiatry*. 2019;27:809–19.
98. O'Brien JT, Holmes C, Jones M, Jones R, Livingston G, McKeith I, Mittler P, Passmore P, Ritchie C, Robinson L, Sampson EL, Taylor JP, Thomas A, Burns A. Clinical practice with anti-dementia drugs: A revised (third) consensus statement from the British Association for Psychopharmacology. *Journal of Psychopharmacology*. 2017;31:147–68.
99. Budd Haeberlein S, Aisen PS, Barkhof F, Chalkias S, Chen T, Cohen S, Dent G, Hansson O, Harrison K, von Hehn C, Iwatsubo T, Mallinckrodt C, Mummery CJ, Muralidharan KK, Nestorov I, Nisenbaum L, Rajagovindan R, Skordos L, Tian Y, van Dyck CH, Vellas B, Wu S, Zhu Y, Sandrock A. Two Randomized Phase 3 Studies of Aducanumab in Early Alzheimer's Disease. *Journal of Prevention of Alzheimer's Disease*. 2022;9:197–210.
100. van Dick C, Swanson C, Aisen P, Bateman R, Chen C, Gee M, Kanekijo M, Reyderman L, Li D, Cohen S, Froelich L, Katayama S, Sabbagh M, Watson D, Dhadda S, Irizarry M, Kramer L, Iwatsubo T. Lecanemab in Early Alzheimer's Disease. *N Engl J Med*. 2023;338:9–21.

Table 1 Epidemiological evidence on prevalence of dementia in patients with heart failure

Study	Sample size of HF patients (N)	Mean age	Prevalence of cognitive impairment	Cognitive impairment definition
Ahluwalia, 2012 (USA)	18 332	81 (75–86)	21.3%	CCW
Brunén, 2021 (ES)	3845	79 ± 9	16%	SPMSQ
Biagi, 2014 (IT)	1444	79 ± 9	35%	Pfeiffer test
Chitnis, 2015 (USA)	8062	74 ± 9	14.1%	ICD-9
Cancian, 2013 (IT)	1905	78 (65–80)	15.7%	GP data sheet
Carter, 2016 (UK)	31 760	74 ± 14	6.8%	ICD-10
Fujiki, 2022 (JPN)	1852	76 ± 8	6.9%	Previous medical reports
Yaku, 2018 (JPN)	3555	80 (70–90)	16.8%	Physician records
Lafo, 2022 (USA)	21 655	77 ± 10	36%	ICD-9
Lee, 2019 (International)	1846	61 ± 11	13.6%	≥2 MMSE points drop
Frey, 2018 (GER)	148	65 ± 10	41%	Test Battery of Attentional Performance
Sterling, 2019 (USA)	436	70 ± 9	14.9%	SIS
Sun, 2017 (CAN)	90 707	73 ± 13	3.2%	ICD-10

Abbreviations: CCW, Chronic Condition Warehouse; GP, general practitioner; ICD, International Classification of Disease; MMSE, Mini-Mental State-Examination; SIS, Six-Item Screener; SPMSQ, Short Portable Mental Status Questionnaire.

Table 2 Epidemiological evidence on incidence of dementia in patients with heart failure

Study	Sample size (N)	HF prevalence	Mean age	Cumulative incidence (%)/hazard ratio of cognitive impairment in HF patients (follow-up duration)	Cognitive impairment definition
Aldeborg, 2017 (DK)	1 946 497	16.7%	77 (69–84)	1.21 (95% CI 1.18–1.24) (6 years)	ICD
Bressler, 2017 (USA)	9895	12.4%	59 ± 6	1.43 (95% CI 1.24–1.66)	DSST
				1.43 (95% CI 1.24–1.66)	DWRT
				1.12 (95% CI 0.98–1.28) (6 years)	WFT
Cacciatore, 1998 (IT)	1075	8%	75 ± 7	1.96 (95% CI 1.07–3.58)	MMSE
de Bruijn, 2015 (NL)	7003	3%	69 ± 9	0.87 (95% CI 0.59–1.28) (8 years)	DSM-III
Hammond, 2018 (USA)	4864	10.2%	76 ± 6	1.70 (95% CI 1.2–2.2) (5 years)	MMSE; DSST
Haring, 2013 (US)	6455	1%	73 (60–84)	1.49 (95% CI 0.73–3.04) (8.4 years)	MMSE, neurocognitive and neuropsychiatric objective exam
Legdeur, 2019 (NL)	442 428	6.8%	87 (65–115)	1.26 (95% CI 1.06–1.49) (3.6 years)	GP records
Noale, 2013 (IT)	2501	5%	71 ± 5	1.40 (95% CI 0.5724–3.4532) (7.8 years)	MMSE; ICD-10; DSM III-R
					NINCDS-ADRDA; CAMDEX;
					Pfeiffer test;
					neurological examination
Qiu, 2006 (SW)	1301	15%	83 ± 5	1.70 (95% CI 1.24–2.34) (5 years)	MMSE
Peters, 2009 (UK)	3336	0.4%	>80	0.55 (95% CI 0.08–3.91) (1.8 years)	MMSE <24 or drop ≥3 point
					DSM-IV; CERAD
Rusanen, 2014 (FI)	1510	6%	50 ± 6	1.40 (95% CI 1.01–1.93) (25 years)	DSM-IV; NINCDS-ADRDA
				0.87 (95% CI 0.35–2.16) (8 years)	
Wändell, 2018 (SW)	537 513	39%	77 ± 9	0.79 (95% CI 0.76–0.84) (4 years)	ICD-10
Witt, 2018 (USA)	6495	14.7%	76 ± 5	1.60 (95% CI 1.13–2.25) (15 years)	CDR; WMS-III, WAIS-R; DSB, BNT, MMSE

Abbreviations: BNT, Boston Naming Test; CAMDEX, Cambridge Mental Disorders of the Elderly Examination; CDR, Clinical Dementia Rating; CERAD, Consortium to Establish a Registry for Alzheimer's Disease; DSM, Diagnostic and Statistical Manual of Mental Disorders; DSST, Digit Symbol Substitution Test; DWRT, Delayed Word Recall Test; ICD, International Classification of Disease; MMSE, Mini-Mental State Examination; NINCDS-ADRDA, National Institute of Neurological Communicative Disorders and Stroke – Alzheimer's Disease and Related Disorders Association; WAIS-R DSB, Digits Span Backward of Wechsler Adult Intelligence Scale-Revised; WFT, Word Fluency Test; WMS-III, Wechsler Memory Scale-III.

Connecting the dots: A narrative review of the relationship between heart failure and cognitive impairment

Mauro Massucci^{1*}, Maria Giulia Bellicini², Marianna Adamo², Andrea Pilotto^{3,4,5}, Marco Metra², Alessandro Padovani^{3,4,5,6} and Riccardo Proietti⁷

¹Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, Brescia, Italy; ²Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, Brescia, Italy; ³Department of Continuity of Care and Frailty, Neurology Unit, ASST Spedali Civili Brescia Hospital, Brescia, Italy; ⁴Department of Clinical and Experimental Sciences, Neurology Unit, University of Brescia, Brescia, Italy; ⁵Laboratory of Digital Neurology and Biosensors, University of Brescia, Brescia, Italy; ⁶Brain Health Center, University of Brescia, Brescia, Italy; and ⁷Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Chest and Heart Hospital, Liverpool, UK

Abstract

Large clinical data underscore that heart failure is independently associated to an increased risk of negative cognitive outcome and dementia. Emerging evidence suggests that cerebral hypoperfusion, stemming from reduced cardiac output and vascular pathology, may contribute to the largely overlapping vascular dementia and Alzheimer's disease. Despite these insights, cognitive outcomes remain largely overlooked in heart failure management. This narrative review outlines the prevalence and risk of cognitive impairment in heart failure patients, exploring potential shared pathophysiological mechanisms and examining the impact of heart failure therapy on cognitive deficits. Additionally, it discusses clinical implications and suggests future treatment approaches targeting therapeutic outcomes. Cognitive impairment is prevalent among individuals with heart failure, with reported rates varying widely depending on assessment methods. Shared pathological pathways and risk factors, including atrial fibrillation (AF), hypertension, obesity and type 2 diabetes mellitus, suggest a causal link. Mechanisms such as poor perfusion, microembolic events, ischaemic syndromes and cerebral inflammation contribute to this relationship. Moreover, heart failure itself may exacerbate cognitive dysfunction. This emerging understanding posits that vascular dementia and Alzheimer's disease may represent a pathophysiological continuum, driven by both the accumulation of misfolded proteins and cerebrovascular pathology due to cardiovascular dysfunction. Understanding these links is crucial for developing effective treatment strategies. The complex interplay between heart failure and cognitive impairment underscores the necessity for a holistic patient care approach. Both conditions share analogous disease processes, influencing self-management and independence in patients. Prioritizing brain health in heart failure management is essential to enhance patient prognosis and general well-being.

Keywords Brain health; Cognitive impairment; Heart failure

Received: 15 May 2024; Revised: 3 September 2024; Accepted: 11 October 2024

*Correspondence to: Mauro Massucci, Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, Piazzale Spedali Civili, 1, 25123 Brescia, Italy. Email: mauro.massucci@gmail.com

Introduction

Heart failure (HF) is a prevalent condition, affecting at least 26 million people worldwide and posing a significant social and economic impact on public health. Based on a US investigation, healthcare costs for HF vary from \$441 to \$1585 per-patient-per-month (PPPM), depending on the patient's

condition, whether stable or experiencing a worsening HF event.¹ This burden is expected to increase as the global population ages, with HF prevalence predicted to double within 40 years.² Similarly, mild cognitive impairment (CI), defined as a statistical construct denoting performance on cognitive tests consistently below age- and education-specific norms,³ and its severe form, dementia, are also associated with

4.3 Data Specification and AI Grounding: Methodological framework

The Dorian Gray Project's methodology is founded on a rigorous, multi-phased approach, conceptualized to translate the complex causality of the heart-brain axis into an actionable digital twin. This framework is characterized by the establishment of a large-scale multimodal data foundation, the design of a cascading AI model suite, and the integration of a real-time behavioral intervention system.

1. *Multimodal Data Foundation and Harmonization*

The project is grounded in high-volume, heterogeneous retrospective data, a critical step that dictates the feasibility and training of all downstream AI models.

Data Scale and Scope: The foundation leverages data on over 55,000 individuals from 10 international observational and clinical cohorts across four European countries and the USA, with a total of ~235,000 individuals when including real-world clinical data from the USA. The cohorts cover middle-to-older age adults with up to 23 years of follow-up on cognitive function (**Table 1**).

Multimodal Input Specification: The framework mandates the integration of diverse data types to capture the complexity of the disease:

- **Neuroimaging Features (MRI):** Extraction of standardized features, including volumetric indices (e.g., cortical thickness via FreeSurfer), cerebral Small Vessel Disease (cSVD) features (e.g., automated detection of White Matter Hyperintensities and cerebral microbleeds), and hemodynamic/functional connectivity parameters.
- **Omics Data:** Harmonization of metabolomic (up to 251 metabolites) and proteomic data (up to 3072 proteins via Olink/SOMAscan platforms). Harmonization techniques, such as z-scores for relative measurement units and Mendelian Randomization for validation, are used for robust biomarker selection.

Data Architecture (CDM): All heterogeneous data will be harmonized into a Common Data Model (CDM), which is interlinked with international standards like SNOMED-CT, ICD, and OMOP to ensure interoperability and reuse (FAIR principles).

2. *Cascading AI Model Suite (PROBAI Models)*

The AI modeling approach unfolds across three distinct phases, moving from interpretable causality to advanced "black box" prediction, ultimately integrating a language model for knowledge refinement.

Phase 1: Explainable and Causal Modeling (PROBAI_STATIC & PROBAI_DYN): This phase employs interpretable "white box" methods like Random Forest and XGBoost for initial multimodality data exploration and feature selection. Causal Inference is integrated using Causal Bayesian Networks (CBNs) and Fuzzy Logic Probability Trees (FPDT) for static modeling and Dynamic

Table 1. Datasets for development and validation of the AI model

	UK Biobank	Whitehall II	ELSA Study	EPIC-Norfolk Study	Maastricht Study	Republic Kardioviz Study	LIFE-Bio cohort	LIFE Study	BeLOVE	Mark VCID	Mount Sinai Data
Patients (n)	25900	6629	10109	8000	3000	2123	360	1800	365	660	236905
Follow-up	7 years	20 years	15 years	13 years	8 years	7 years	1 year	7 years	2 years	2 years	23 years
Age range (y)	40-69	35-55	50-90	40-75	40-75	25-64	50-90	18-79	18-90	26-90	50-90
Sex	M & F	M & F	M & F	M & F	M & F	M & F	M & F	M & F	M & F	M & F	M & F
Cognition measurement	Multidom Composite	Multidom Composite	Multidom Composite	Multidom Composite	Multidom Composite	MOCA	MOCA	Multidom Composite	MOCA	MOCA	MOCA
Genetics	N=20000	N=5666	N=6185	N=6000	N=3000	-	N=360	N=7669	N=200	-	N=5000
Metabolomics single	Nightingale Health N=251 biomarkers. N=15365	Nightingale Health n=251 biomarkers n=5374	-	DiscoveryHD4® platform, Metabolon, Inc. N=940 metabolites N=1503	Nightingale Health N=226 biomarkers. N=3441	-	-	LC-MS/MS N=9622	Biocrates Quant 500	-	-
multiple	N=3289	-	-	-	N=1850	-	-	N=1,000	-	-	-
Proteomics: single	Olink 3072 PEA N=2982	Olink panel (n=543) & SOMAscan version 4 assay N=2274	NA	Olink 1536 panel (N=1180), 368 proteins from the CVII, CVIII, Cardiometabolic, inflammation panels (N=822)	NA	NA	SIMOA (7 selected proteins)	Olink CVIII, n=96 proteins (N=2016)	OLINK-3368 panel (N=70)	NA	NA
MRI: single	N=14195	N=800	-	-	N=3000	-	360	N=1200	N=50	N=660	-
multiple	N=248	-	-	-	N=2500	-	50	N=1000	N=25	-	-
ECG: single	N=14341	n=5200	-	-	n=3000 (2200 24h ECG) N=3000 (2000 24h ECG)	n=737	360	N=9000	N=365	-	N=200000
multiple	N=250	N=4000	-	-	-	-	-	N=1000	N=25	-	150000

Abbreviations: LC-MS/MS: Liquid chromatography—mass spectrometry and liquid chromatography—tandem mass spectrometry; MR, magnetic resonance imaging. *≥2 cognition measurements and ≥2 CVRF measurements

Bayesian Networks (DBNs) for dynamic modeling, advancing beyond mere correlation to establish causal understanding. This foundation is crucial for interpretability and robustness.

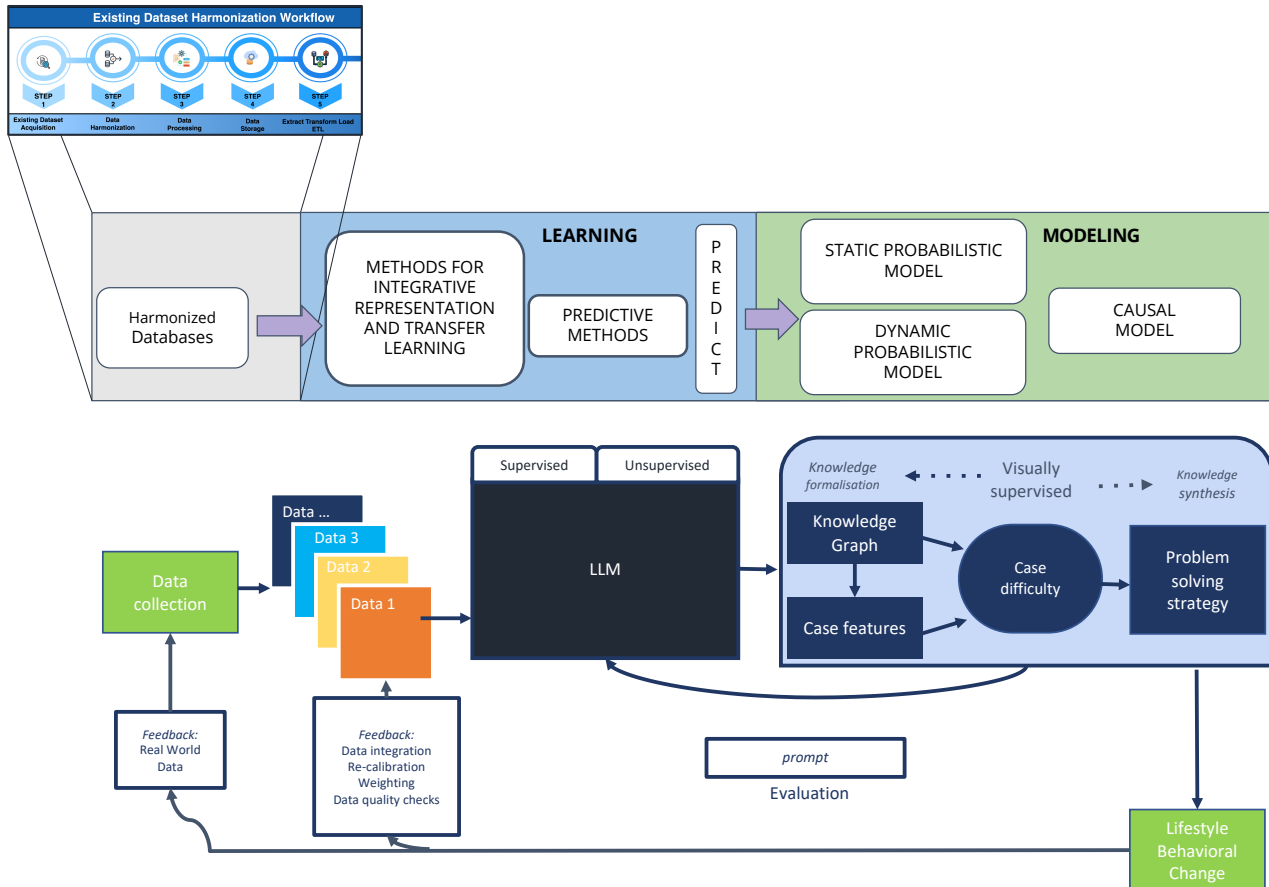
Phase 2: "Black Box" Methods and Digital Biomarkers: The predictive accuracy is enhanced using advanced deep learning techniques, including Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) (or hypergraph neural networks) to process unstructured and relational data. Transformer-based architectures are employed for temporal data analysis, extracting deep causal relationships over time from longitudinal data streams. The core goal is to replace or extend conventional biomarkers with digital features (digital biomarkers) extracted from passive and active data sources (e.g., smartwatches, exergame performance).

Phase 3: Knowledge Refinement (Tiny Language Model): A compact language model (LM) is developed as an add-on and combined with a knowledge graph (generated from causal models) using techniques like Retrieval-Augmented Fine Tuning (RAFT) to specialize the LM for CVD and MCI-related knowledge.

3. Digital Twin & Behavioral Intervention Loop

The final stage involves deploying the AI models into the PORTRAIT platform to support real-time clinical decisions and personalized behavioral change interventions.

The PORTRAIT Platform: This cloud platform acts as the Digital Twin backbone, integrating the AI models and ensuring seamless multi-source data ingestion, scalability, and security.

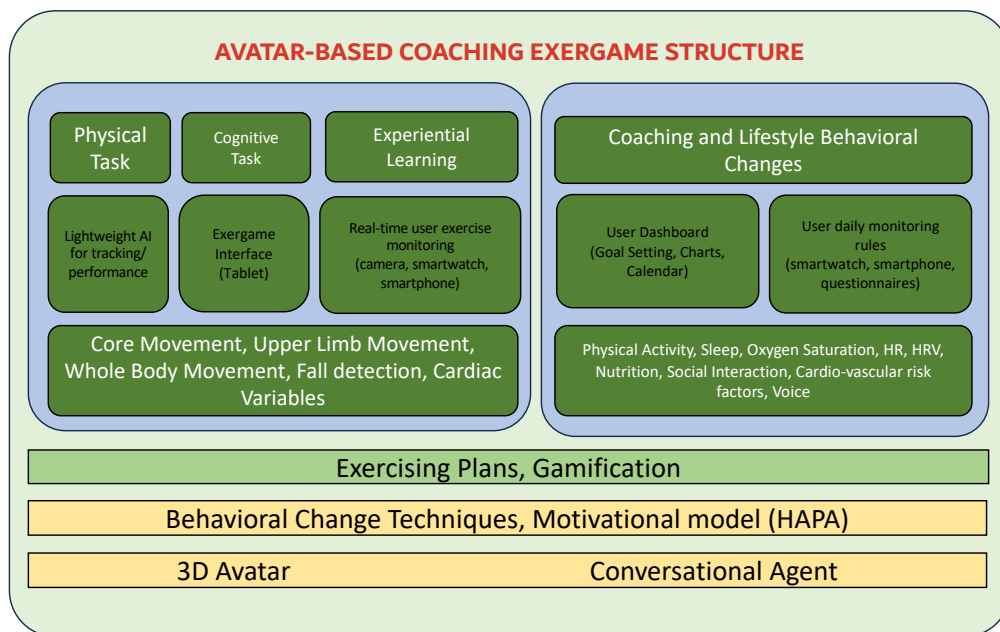


AI-Driven Coaching System (CS): The Digital Twin feeds the CS with real-time predictions and actionable insights. This system implements a Behavioral Change Interventional Framework (ABCE logic), providing personalized guidance and feedback through the MIRROR app.

- The CS is grounded in the Health Action Process Approach (HAPA) model, structuring the intervention into motivational and volitional phases.
- Real-time data from wearables and exergame performance dynamically adapt the physical-cognitive training sessions, ensuring personalization based on adherence and performance levels.

Clinical Dashboard: The system culminates in a dashboard for Healthcare Professionals (HCPs) that integrates all multi-source prospective data, risk predictions, trajectories, and Explainable AI (XAI) versions of the model paths for each patient, facilitating interpretability and informed decision-making.

This comprehensive and layered methodological framework represents a significant and novel design for integrating multimodal AI, causal inference, and dynamic behavioral intervention in cardiovascular health.



4.4. Clinical Trial Design for AI System Validation: The COGIT-HF protocol

As the final, crucial step in validating the DORIAN GRAY Digital Twin and its therapeutic strategy, the project will be tested in the investigator-initiated COGIT-HF trial. This randomized controlled trial (RCT) is designed to evaluate the feasibility and initial efficacy of the AI-augmented multidomain intervention in a high-risk patient population, directly addressing the limitations of current non-specific interventions in the heart-brain axis.

Trial Objectives and Rationale

The trial's design is driven by the urgent need for personalized interventions for cognitive decline in heart failure (HF) patients. Existing multidomain interventions show promise (e.g., FINGER study²⁸), but lack personalization and struggle with adherence. COGIT-HF aims to overcome these limitations using the digital twin approach.

Objective Category	Specific Goal
Primary Objective (Feasibility)	To evaluate the feasibility of the avatar-based coaching exergame (ABCE) on top of Optimal Medical Therapy (OPT) versus OPT alone in improving cognitive function in patients with HF and MCI across five European countries.
Secondary Objectives	To assess adherence to the digital treatment and its impact on quality of life (KCCQ-23 scores).
Exploratory / Translational Objectives	Inform sample-size calculation for a future full-scale RCT, develop and refine the patient-specific digital-twin model using longitudinal data, and identify and validate novel digital biomarkers from continuous monitoring.

Study Design and Population

The COGIT-HF trial is structured as a pilot randomized controlled trial.

Trial Setting: Participants will be recruited from at least five hospitals throughout Europe.

Study Population (Inclusion Criteria):

- Men and women aged ≥ 55 years.
- Chronic Heart Failure (HF) defined according to ESC 2021 guidelines (LVEF $< 50\%$ and/or increased natriuretic peptides).
- Mild Cognitive Impairment (MCI) defined as $18 \leq \text{MoCA score} \leq 26$.

Intervention and Randomization: Eligible participants will be randomized to one of two arms:

- Standard of Care Arm (OPT): Optimal Medical Therapy alone.
- Intervention Arm (OPT + ABCE): Optimal Medical Therapy plus access to the Digital Avatar-Based Coaching Exergame (ABCE).

Randomization is stratified by country and uses random permuted blocks.

The Intervention: Avatar-Based Coaching Exergame (ABCE)

The intervention arm receives the ABCE, which is the core of the digital therapeutic tested in the DORIAN GRAY trial. It delivers a multidomain intervention through the MIRROR app.

Multidomain Focus: The program integrates dual-task paradigms, combining physical exercise (Aerobic/Cardiovascular, Strength, Flexibility, Balance) with cognitive stimulation (Attention, Memory, Problem Solving).

Personalization Loop: Participants engage in structured sessions 3–5 times per week. A smartwatch continuously collects physiological data (e.g., heart rate, sleep, activity). This real-time data feeds the digital twin model to continuously adapt and personalize the exercise intensity, difficulty, and cognitive task complexity, ensuring safety and optimal progression.

Coaching Component: The digital coach provides personalized feedback and supports experiential learning, reinforcing medication adherence and healthy lifestyle behaviors.

Key Outcome Measures

The trial measures outcomes reflecting both clinical status and the technological success of the digital twin design.

Outcome	Measurement Timepoint	Focus
Primary Endpoint	Change in NTB Global Z-score (a composite cognitive measure)	Clinical Efficacy/Cognition
Key Secondary Endpoints	Changes in CANTAB remote-testing, Quality of Life (KCCQ scores), and adherence to the digital treatment	Quality of Life/Adherence
Mechanistic Endpoints	Changes in Blood Biomarkers (pTau217, NfL, GFAP), and Functional MRI (resting state imaging) at baseline and 12 months	Biological Mechanism
Digital Biomarkers (Exploratory)	Assessing predictive performance of combined models (clinical + biological + digital) using continuous data features (e.g., HRV, sleep fragmentation, in-app accuracy/time-to-response metrics)	AI Validation/Model Refinement

The COGIT-HF trial is critical to fulfilling the translational aims of the doctorate, demonstrating the viability of using an AI-augmented digital twin to deliver personalized, effective care to complex, high-risk cardiac populations.

5. Conclusions

The goal of this doctoral work, "Applications of Artificial Intelligence in Cardiology: Methodological Advances and Three Use Cases," was to demonstrate the transformative potential of artificial intelligence in cardiology, shifting the paradigm from reactive to predictive, personalized, and proactive healthcare. This thesis successfully explored three major, complementary applications of AI, each addressing a critical unmet clinical need and contributing significantly to the methodological landscape of AI in Medicine.

5.1. Main Contributions of This PhD

The collective contribution of this thesis is defined by the strategic application of diverse AI technologies across three distinct levels of the clinical ecosystem: Opportunistic Screening, Foundation Models, and Personalized Digital Intervention.

1. Contribution to Scalable and Low-Cost Screening (Use Case 1)

This work established the viability of Opportunistic Screening by successfully extracting latent prognostic value from ubiquitous, low-cost imaging.

Clinical Impact: We demonstrated that Deep Learning (DL) models applied to routine Chest X-Rays (CXR) can accurately predict significant Coronary Artery Disease (CAD) for diagnostic triage and long-term all-cause mortality in high-risk cohorts.

Methodological Innovation: The research confirmed that the DL model could differentiate the severity and extension of CAD and that the DL-derived CXR risk score acts as a strong independent determinant of severe CAD and long-term mortality, often outperforming traditional clinical risk scores like the Diamond-Forrester score. This validates CXR as a powerful source of underutilized prognostic biomarkers.

2. Contribution to Generalizable Diagnostic Tools (Use Case 2)

This work marked a methodological leap forward by pioneering the Foundation Model paradigm for ECG interpretation, addressing the inherent limitations of fragmented, task-specific AI systems.

Clinical Impact: We developed HuBERT-ECG, a large-scale, self-supervised foundation model pre-trained on 9.1 million ECGs. This model proved highly versatile and generalized, achieving high-accuracy detection for 164 cardiovascular conditions and demonstrating capability in three major roles: primary diagnosis, supportive diagnosis, and predicting future cardiovascular events.

Methodological Innovation: By using self-supervised learning on vast, unlabeled data, HuBERT-ECG significantly enhanced performance, particularly in complex scenarios (e.g., boosting AUROCs for

Heart Failure and Coronary Heart Disease prediction when fine-tuned on multimodal data). This establishes a new, flexible baseline for AI-ECG applications.

3. Contribution to Personalized Digital Intervention (Use Case 3)

This work provided the foundational blueprint for constructing an AI-augmented Digital Twin ecosystem, laying the methodological and clinical ground for next-generation personalized prevention.

Clinical Impact: Through the Horizon Europe-funded DORIAN GRAY Project, we conceptualized a multimodal, personalized, and adaptive intervention to target the critical link between Cardiovascular Disease and Mild Cognitive Impairment.

Methodological Innovation: We designed a rigorous framework for integrating high-dimensional retrospective data (Omics, Neuroimaging) with Real-World Data (RWD) from smartwatches via the PORTRAIT platform. This enables a cascading AI model suite (PROBAI Models) focused on causal inference and dynamic prediction, validated through the dedicated COGIT-HF randomized trial protocol.

5.2. Broader Implications and Recommendations

The findings across these three use cases offer clear implications for both clinical practice and public health stakeholders:

- For the Healthcare System (Efficiency and Equity): The demonstrable power of AI in repurposing ubiquitous and affordable tools like CXR and ECG validates a strategy for large-scale, low-cost screening. Deploying such systems can enhance efficiency, rationalize diagnostic triage, and promote more equitable access to preventative risk stratification.
- For Clinicians (Prediction and Decision Support): Foundation models like HuBERT-ECG represent a paradigm shift from narrow, single-task algorithms to versatile, generalizable AI allies that can extract multi-dimensional physiological fingerprints, thereby enhancing precision and broadening the diagnostic capability of routine tests.
- For Patients (Personalization and Adherence): The Digital Twin methodology developed in DORIAN GRAY offers a path toward truly personalized medicine. By dynamically adapting interventions (ABCE) based on continuous individual data, AI can overcome traditional barriers to adherence, fostering patient empowerment and enhancing the efficacy of multidomain therapies.
- For Future Research (Methodological Advancements): The challenges encountered underscore the ongoing need for research into Explainable AI (XAI), robust model validation in diverse

cohorts, and the ethical management of complex, real-world data streams to build trust and ensure fairness in AI-assisted medicine.

In conclusion, this thesis confirms that Artificial Intelligence is not merely an auxiliary tool but a foundational element of future cardiovascular medicine, capable of delivering measurable advancements in risk prediction, patient management, and health system efficiency.

References

1. Maddox, T. M. *et al.* Generative AI in Medicine - Evaluating Progress and Challenges. *N Engl J Med* **392**, 2479–2483 (2025).
2. Ning, Y. *et al.* Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *Lancet Digit Health* **6**, e848–e856 (2024).
3. Topol, E. J. AI-enabled opportunistic medical scan interpretation. *Lancet* **403**, 1842 (2024).
4. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision-language foundation model for echocardiogram interpretation. *Nat Med* **30**, 1481–1488 (2024).
5. Thangaraj, P. M., Benson, S. H., Oikonomou, E. K., Asselbergs, F. W. & Khera, R. Cardiovascular care with digital twin technology in the era of generative artificial intelligence. *Eur Heart J* **45**, 4808–4821 (2024).
6. Roth, G. A. *et al.* Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol* **76**, 2982–3021 (2020).
7. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. *Eur Heart J* **32**, 1316–1330 (2011).
8. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* **357**, j2099 (2017).
9. Vrints, C. *et al.* 2024 ESC Guidelines for the management of chronic coronary syndromes. *Eur Heart J* **45**, 3415–3537 (2024).
10. Lin, A. *et al.* Artificial Intelligence in Cardiovascular Imaging for Risk Stratification in Coronary Artery Disease. *Radiol Cardiothorac Imaging* **3**, e200512 (2021).
11. Lin, A. *et al.* Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study. *Lancet Digit Health* **4**, e256–e265 (2022).
12. Yoon, S.-H. *et al.* Improving Image Quality of Chest Radiography with Artificial Intelligence-Supported Dual-Energy X-Ray Imaging System: An Observer Preference Study in Healthy Volunteers. *J Clin Med* **14**, (2025).
13. Klein, J. S. & Rosado-de-Christenson, M. L. A Systematic Approach to Chest Radiographic Analysis. in *Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging* (eds. Hodler, J., Kubik-Huch, R. A. & von Schulthess, G. K.) 1–16 (Springer, Cham (CH), 2019). doi:10.1007/978-3-030-11149-6_1.
14. Farina, J. M. *et al.* Artificial Intelligence-Based Prediction of Cardiovascular Diseases from Chest Radiography. *J Imaging* **9**, (2023).
15. Parsa, S., Shah, P., Doijad, R. & Rodriguez, F. Artificial Intelligence in Ischemic Heart Disease Prevention. *Curr Cardiol Rep* **27**, 44 (2025).

16. Lu, M. T. *et al.* Deep Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA Netw Open* **2**, e197416 (2019).
17. Bader, A. S., Rubinowitz, A. N., Gange, C. P. J., Bader, E. M. & Cortopassi, I. O. Imaging in the Evaluation of Chest Pain in the Primary Care Setting, Part 1: Cardiovascular Etiologies. *The American journal of medicine* **133**, 1033–1038 (2020).
18. Lu, M. T. *et al.* Deep Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA network open* **2**, e197416 (2019).
19. Cook, D. A., Oh, S.-Y. & Pusic, M. V. Accuracy of Physicians' Electrocardiogram Interpretations: A Systematic Review and Meta-analysis. *JAMA Intern Med* **180**, 1461–1471 (2020).
20. Kalmady, S. V. *et al.* Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. *NPJ Digit Med* **7**, 133 (2024).
21. Attia, Z. I., Harmon, D. M., Behr, E. R. & Friedman, P. A. Application of artificial intelligence to the electrocardiogram. *Eur Heart J* **42**, 4717–4730 (2021).
22. Wouters, P. C. *et al.* Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. *Eur Heart J* **44**, 680–692 (2023).
23. Li, J. *et al.* An Electrocardiogram Foundation Model Built on over 10 Million Recordings. *NEJM AI* **2**, (2025).
24. Hughes, J. W. *et al.* A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *NPJ Digit Med* **6**, 169 (2023).
25. Sadée, C. *et al.* Medical digital twins: enabling precision medicine and medical artificial intelligence. *Lancet Digit Health* **7**, 100864 (2025).
26. Proietti, R., Adamo, M. & Massucci, M. The DORIAN GRAY project: bridging cardiovascular disease and cognitive decline. *Eur Heart J* **46**, 4456–4458 (2025).
27. Saeed, A., Lopez, O., Cohen, A. & Reis, S. E. Cardiovascular Disease and Alzheimer's Disease: The Heart-Brain Axis. *J Am Heart Assoc* **12**, e030780 (2023).
28. Kivipelto, M. *et al.* The Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER): study design and progress. *Alzheimers Dement* **9**, 657–665 (2013).