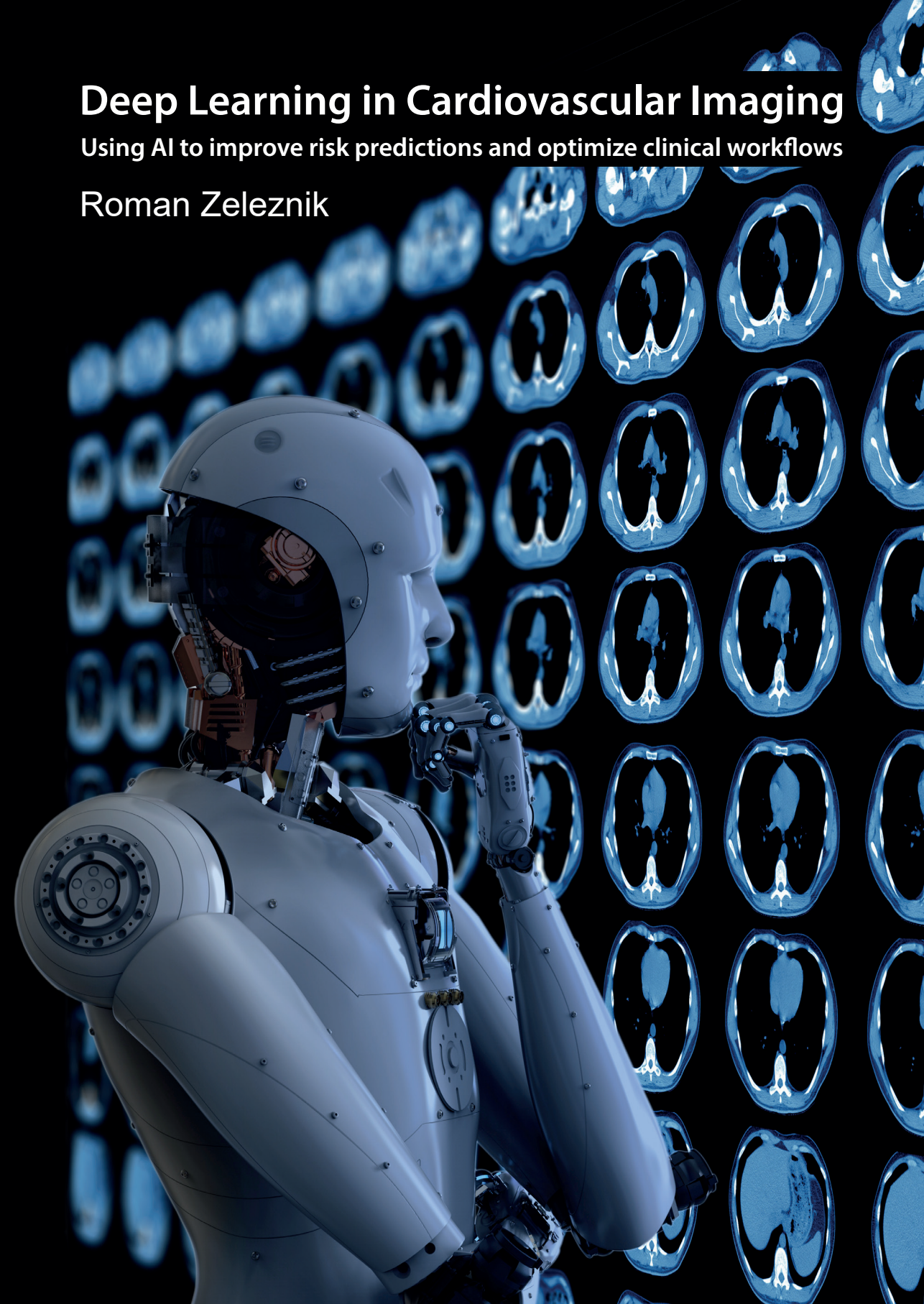
Deep learning in cardiovascular imaging

Deep Learning in Cardiovascular Imaging

Using AI to improve risk predictions and optimize clinical workflows

Roman Zeleznik



Deep Learning in Cardiovascular Imaging

Using AI to improve risk predictions and optimize clinical workflows

Roman Zeleznik

Promotors
Prof. Dr. Ir. Hugo Aerts
Prof. Dr. Udo Hoffmann (Harvard University, Boston, USA)

Assessment committee
Prof. Dr. Ir. Andre Dekker, Chairman
Prof. Dr. Bram van Ginneken (Radboudumc, Nijmegen)
Prof. Dr. M. Eline Kooi
Prof. Dr. Casper Mihl
Prof. Dr. Wiro J. Niessen (Erasmus Medisch Centrum, Rotterdam)

Deep Learning in Cardiovascular Imaging

Using AI to improve risk predictions and optimize clinical workflows

Dissertation

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus Prof.dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public on
Thursday, 16th of September 2021 at 14:00 hrs.

by

Roman Zeleznik

Roman Zeleznik
Deep Learning in Cardiovascular Imaging
Using AI to improve risk predictions and optimize clinical workflows

PhD thesis, Maastricht University, Maastricht, the Netherlands (2021)

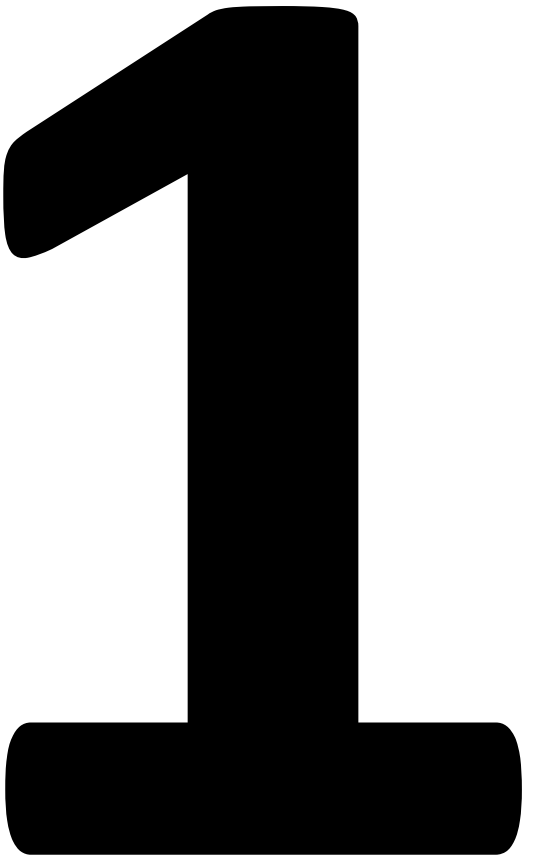
ISBN: 978-94-6423-409-1

The research described in this thesis was financially supported by National Institutes of Health; European Research Council; German Research Foundation; National Heart, Lung and Blood Institute; American Heart Association; Fulbright Visiting Researcher Grant; Rosztoczy Foundation Grant.

Financial support from Maastricht University for printing this thesis is gratefully acknowledged.

Cover design: Graphics design: Roman Zeleznik
Droid: Phonlamai Photo - Shutterstock
Medical images: Kalewa - Shutterstock
Cover layout: Mark van 't Veer
Layout: Dennis Hendriks || ProefschriftMaken.nl
Printing: ProefschriftMaken.nl

Table of Content



Chapter 1

General introduction and outline of the thesis

Introduction

This thesis aims to develop deep learning methods for cardiac segmentation and prediction tasks in computed tomography (CT) to investigate heart characteristics and their predictive value for future cardiac events, support medical experts in daily routine tasks and accelerate and improve clinical treatment. This section will provide an introduction to machine learning, present different concepts and applications, followed by the thesis outline.

Deep Learning in medical imaging

Deep learning (DL) is a sub-field of machine learning that has made huge improvements in recent years, utilizing the high computational power of state-of-the-art computer hardware and the availability of massive amounts of digital data, achieving equal and even superior performance to humans in task-specific applications[1–4]. Typically, traditional machine learning systems consist of one or two transformation layers[5–7]. DL emerged from the idea of stacking multiple layers in a ML system, with "deep" describing the number of layers as the depth of a system[8–10], that has shown great results in many areas (e.g. face recognition[11], speech recognition[12], natural language processing[13], self driving cars[14], playing AlphaGo[4], and many more). Furthermore, DL was successfully developed and applied in several medical applications, such as medical imaging and diagnostic, risk management, or virtual assistants, as well as medical image segmentation tasks, with U-Nets and U-Net-derived networks being prominent models[15–19]. These networks were successfully applied for brain tumor segmentation[20], lung nodule segmentation[21], whole heart segmentation[22,23], and others. Recently, the success of deep learning has led to a vast amount of proof-of-principle studies in various clinical application areas[24–26]. However, often real world applicability of the proposed systems was not demonstrated due to a lack in sufficiently large, diverse and independent data sets for model development and validation. Hence, before clinical introduction can be considered, generalizability of these systems needs to be demonstrated as they need to perform reliably across multiple clinical scenarios, and work robustly on data recorded with various settings across institutions.

In this thesis, novel deep learning methods for clinical applications were developed and evaluated. Their performance, robustness, and generalizability was rigorously assessed in large, independent and distinctive cohorts. The presented deep learning methods focus on enhancing, supporting and improving clinical treatment of cardiac diseases.

Deep learning for cardiovascular risk prediction

Cardiovascular disease is the most common preventable cause of death, accounting for up to 45% of mortality in Europe[27] and 31% in the United States[28]. One of the strongest known predictors of adverse cardiovascular events is coronary artery calcification, which

can be quantified on computed tomography (CT)[29,30]. The CT coronary calcium score is a measure of the burden of coronary atherosclerosis and is one of the most widely accepted measures of cardiovascular risk[29,30]. Coronary calcium scoring has been recommended by guidelines for risk stratification, specifically in the setting of primary prevention in asymptomatic individuals[31,32]. In symptomatic patients, the presence of coronary calcium is associated with future cardiovascular events in the stable chest pain setting[33] and low likelihood of acute coronary syndrome in patients with acute chest pain[34]. Additionally, showing patients their coronary calcium provides a "teachable moment" to empower them to make informed, individualized decisions, and to improve long-term compliance for preventative therapy and lifestyle changes including smoking cessation[35,36].

While the calcium score has been traditionally measured on specialized ECG-gated cardiac CT, it can also be measured on nearly every standard chest CT, performed without contrast agent[32]. However, the measurement requires radiological expertise, time, and specialized coronary calcium quantification equipment. As a result, this essentially free available information is usually not reported. A deep learning system for fully automatic coronary calcium quantification could help put this actionable information into the hands of patients and their physicians.

Deep learning for whole heart segmentation in computed tomography scans

Whole heart segmentation in CT is a critical task performed in medical research and care[22,23]. Moreover, heart segmentation represents the foundation for a wide range of applications, such as coronary calcium segmentation[37] or quantification of pericardial fat[38]. According to the most recent European Society of Cardiology Guidelines, cardiac CT represents a first-line diagnostic method to assess cardiovascular risk in patients with chronic coronary syndromes, including those with stable chest pain, [39]. It is increasingly used to exclude obstructive coronary artery disease in patients presenting with stable chest pain. Still, assessment of cardiovascular risk in CT remains difficult, especially in those with non-obstructive diseases[40], which account for the majority of future cardiovascular events[41,42], thus requiring advanced risk stratification and frequently are referred to further testing. On the other hand, cardiac CT has the advantage to image anatomical structures. For example, the diameter of the heart is an established predictor of cardiovascular risk, which is traditionally measured on X-ray[23,43,44]. However, the prognostic value of CT-derived whole heart volume, a detailed 3D measure of heart size, available in all cardiac CT scans, has not been evaluated yet. Additionally, epicardial adipose tissue, located within the pericardial sac and neighboring coronary arteries, represents a metabolically active tissue with paracrine atherogenic effects related to hypertension, dyslipidemia, diabetes mellitus, and obesity[45,46]. There is a growing body of evidence, mainly based on data from large cohorts with low cardiovascular risk, suggesting that elevated epicardial adipose tissue volume may be associated with coronary artery disease severity and adverse cardiac events[47]. However, little is known about the predictive value of epicardial adipose tissue

in symptomatic patients with elevated cardiovascular risk. Therefore, a publicly available, robust and fully automatic heart segmentation system has the potential to accelerate cardiac research and could allow translation to medical care.

Deep learning systems beyond the initial field of development

Medical knowledge is increasing exponentially with an estimated doubling every few months as of 2020[48]. While this has improved healthcare across the world[49], it is paralleled by increasingly specialized expert knowledge, which may be disproportionately distributed to high resource medical centers, thus increasing health care disparities[50]. Recent advances in AI, and deep learning in particular, offer a novel way to improve and automate complex tasks that up until now could only be performed by professionals[51]. Typically, deep learning applications are developed using labeled data generated by medical experts for domain-specific problems. As a result, this expert knowledge is encapsulated in the deep learning system, providing the opportunity to disseminate this highly skilled expertise across medical domains, institutions and countries, with the potential to optimize patient care and reduce knowledge and economic disparities in undersupplied settings.

One area that could benefit from this concept includes imaging-related tasks, such as radiology and radiation oncology. While the former uses imaging studies primarily for diagnosis, the latter relies on the same information for organ and tumor targeting, treatment planning and delivery, and monitoring. An integral part of radiotherapy treatment planning is segmenting organs at risk in the radiation field on CT scans[22]. If appropriate resources are available, this is done manually by trained experts who require considerable time and are prone to inter- and intra-observer variability. More importantly, if time or knowledge are limited, this crucial step to ensure treatment quality and patient safety may be neglected. Therefore, automating and optimizing the process of organ at risk segmentation using deep learning could improve clinical care at high speed and low additional cost, especially in underprivileged healthcare settings[52].

Among the organs at risk, the heart is of special interest as it is known that increasing radiation exposure is associated with future cardiac adverse events, such as coronary artery disease and heart failure[53,54]. Given their training, the highest anatomic expertise in cardiac imaging is likely found among cardiovascular radiologists, who focus on the diagnosis and monitoring cardiac-related diseases using dedicated image acquisition, reconstruction, and analysis techniques. Hence, disseminating this highly specific but narrow expert knowledge across medical domains and to institutions or countries with limited resources may enable more accurate treatment planning and measurement of cardiac radiation dose to optimize cardioprotective strategies in radiation oncology. This is of particular interest for patients with breast cancer as the heart and its substructures are in close proximity to the target area. Thus, reducing heart radiation dose is of great importance to not harm the generally favourable outcomes of these patients.

Objectives and outline of the thesis

To address the mentioned challenges, several deep learning models for cardiac related tasks were developed in this thesis. In several projects a cooperation between technical and medical experts from renowned scientific and medical institutes was formed to process and analyze data from unprecedentedly large clinical cohorts.

Chapter 2 presents a DL system for automatic cardiovascular risk prediction. The proposed system consists of four consecutive steps (1) to localize and (2) segment the heart in a CT scan, (3) segment coronary calcium and (4) calculate a risk score. The system was developed using cardiac ECG-gated CT scans from the community based, observational Framingham Heart Study (FHS)[55]. It was tested in over 20,000 CT scans from 4 distinct clinical cohorts, including low-dose chest screening CTs of asymptomatic individuals from the National Lung Screening Trial (NLST)[56] as well as cardiac gated CTs of individuals with acute chest pain from the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE)[57] and Rule Out Myocardial Infarction using Computer Assisted Tomography II study (ROMICAT-II)[58]. The scans were gathered by over 200 participating medical sites and manual segmentations were provided by experienced medical experts from the Cardiovascular Imaging Research Center at the Massachusetts General Hospital. Accuracy, compared to the gold standard of expert human readers, was assessed in 5,521 subjects across all four cohorts. The results demonstrate that deep learning methods can automate cardiovascular risk prediction from medical images acquired in several clinical scenarios. These observations provide a rationale to implement this technique in both screening and hospital settings to improve population health, at high speed and low costs.

Chapter 3 presents a reliable and accurate deep learning system for fully automatic heart segmentation in non-contrast enhanced cardiac ECG-gated CT and low-dose chest screening CT scans. The model was trained on high quality scans from FHS[59,60], with manual segmentations provided by experienced medical experts. The system was tested on two large and distinct clinical cohorts including cardiac ECG-gated CTs from PROMISE[57] as well as low-dose chest screening CTs from NLST[56]. In the test cohorts totalling 1,534 CT scans acquired in 226 medical sites, performance, generalizability, and applicability of the proposed deep learning system was assessed. The system is suitable for a vast array of research and medical applications. By making the code open-source and providing the fully trained model to the public without restrictions, this study has the potential to accelerate clinical research and improve medical treatment.

Chapter 4 presents a study to determine the association of whole heart volume with major adverse cardiovascular events, adjusting for traditional measures of cardiovascular risk and coronary artery disease (CAD) characteristics on CT. Furthermore, a subgroup analysis across CAD categories was performed to determine whether whole heart volume had discriminatory capacity incremental to atherosclerosis cardiovascular disease risk score and CT-derived CAD characteristics.

Chapter 5 investigates the relationship between epicardial adipose tissue volume and traditional cardiovascular risk factors, CT-derived coronary artery disease characteristics, and incident adverse events in symptomatic individuals with increased cardiovascular risk from PROMISE[57].

Chapter 6 investigates whether a deep learning system developed in cardiovascular radiology can be used for radiation oncology treatment planning. Therefore, a deep learning system for whole heart segmentation was developed using multi-center data including dedicated cardiac CTs and low dose chest screening CTs, with manual segmentations from expert cardiovascular radiologists. This system was validated in an independent real-world dataset including 5,677 breast cancer patients which were treated at the Dana-Farber and Brigham and Women's Cancer Center between 2008-2018. The performance of the deep learning system was compared to radiation oncology experts as well as to heart segmentations used in the clinic for treatment delivery. This study may serve as proof of principle to repurpose and leverage AI applications for optimizing patient care and reducing healthcare disparities across specialties, institutions, and countries.

Chapter 7 presents a summary of this work and next steps, future improvements and applications.

2

Chapter 2

Deep convolutional neural networks to predict cardiovascular risk from computed tomography

Roman Zeleznik, Borek Foldyna, Parastou Eslami, Jakob Weiss, Alexander Ivano, Jana Taron, Chintan Parmar, Raza M. Alvi, Dahlia Banerji, Mio Uno, Yasuka Kikuchi, Julia Karady, Lili Zhang, Jan-Erik Scholtz, Thomas Mayrhofer, Asya Lyass, Taylor F. Mahoney, Joseph M. Massaro, Ramachandran S. Vasan, Pamela S. Douglas, Udo Hoffmann*, Michael T. Lu*, Hugo J.W.L. Aerts*

Published in: Nature Communications (2021)

Abstract

Coronary artery calcium is an accurate predictor of cardiovascular events. While it is visible on all computed tomography (CT) scans of the chest, this information is not routinely quantified as it requires expertise, time, and specialized equipment. Here, we show a robust and time-efficient deep learning system to automatically quantify coronary calcium on routine cardiac-gated and non-gated CT. As we evaluate in 20,084 individuals from distinct asymptomatic (Framingham Heart Study, NLST) and stable and acute chest pain (PROMISE, ROMICAT-II) cohorts, the automated score is a strong predictor of cardiovascular events, independent of risk factors (multivariable-adjusted hazard ratios up to 4.3), shows high correlation with manual quantification, and robust test-retest reliability. Our results demonstrate the clinical value of a deep learning system for the automated prediction of cardiovascular events. Implementation into clinical practice would address the unmet need of automating proven imaging biomarkers to guide management and improve population health.

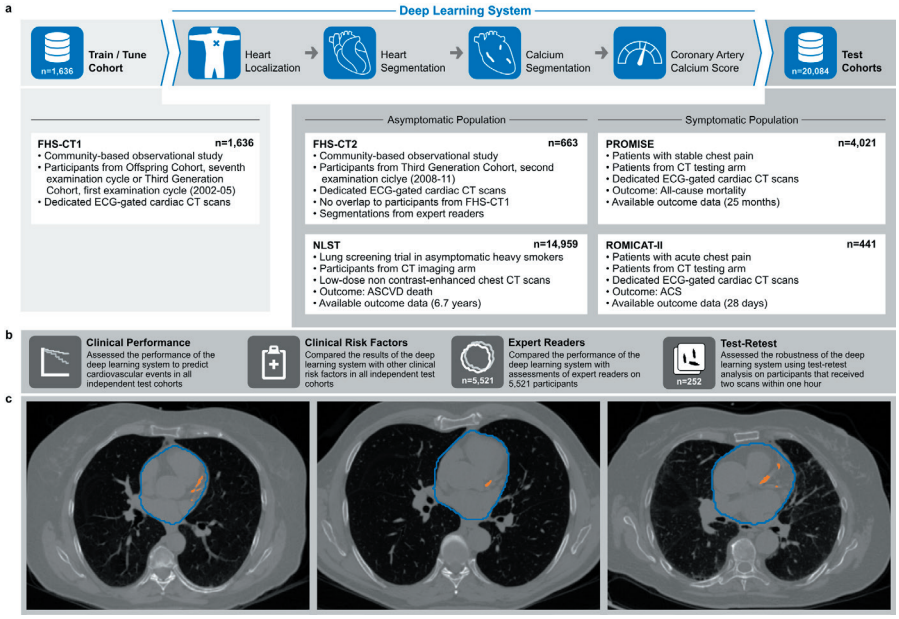


Figure 1. Overview of the deep-learning framework, the training and test cohorts, and the implemented evaluation steps. a The deep-learning framework was trained and tuned on 1,636 computed tomography (CT) scans from FHS-CT1. In four consecutive steps a coronary calcium risk score was calculated in a fully automatic fashion. Independent testing was performed on 20,084 CT scans from four different clinical cohorts. **b** The performance of the framework was evaluated with respect to its clinical value and robustness. **c** CT scans of three representative patients of FHS-CT2 outlined with the deep learning system heart (blue contours) and coronary calcium (orange contours). FHS-CT1[17], FHS-CT2[17]: Framingham Heart Study, (CT1) participants from the seventh examination cycle of the Offspring Cohort or first examination cycle of the Third Generation Cohort (2002-05) and (CT2) participants from the second examination cycle of the Third Generation Cohort (2008-11); NLST[18]: National Lung Screening Trial; PROMISE[19]: Prospective Multicenter Imaging Study for Evaluation of Chest Pain; ROMICAT-II[20]: Rule Out Myocardial Infarction using Computer Assisted Tomography II; ECG: Electrocardiographic; CT: Computed tomography; ASCVD: Atherosclerotic cardiovascular disease; ACS: Acute coronary syndrome.

Introduction

Cardiovascular disease is the most common preventable cause of death, accounting for up to 45% of mortality in Europe[1] and 31% in the United States[2]. Effective lifestyle and pharmacological prevention is available, but identifying those who would benefit most remains an ongoing challenge[3]. Traditional risk factors, such as age and sex, have limited accuracy for predicting cardiovascular desease among individuals. Hence, efforts are needed to further improve cardiovascular risk prediction and stratification on an individual basis[4].

One of the strongest known predictors for adverse cardiovascular events is coronary artery calcification, which can be quantified on computed tomography (CT)[5,6]. The CT coronary calcium score is a measure of the burden of coronary atherosclerosis and is one of the most widely accepted measures of cardiovascular risk[5,6]. Coronary calcium scoring has been recommended by guidelines for risk stratification, specifically in the setting of primary prevention in asymptomatic individuals[7,8]. In symptomatic patients, the presence of coronary calcium is associated with future cardiovascular events in the stable chest pain setting[9] and low likelihood of acute coronary syndrome in patients with acute chest pain[10]. Additionally, showing patients their coronary calcium provides a "teachable moment" to empower them to make informed, individualized decisions, and to improve long-term compliance for preventative therapy and lifestyle changes including smoking cessation[11,12].

While the calcium score has been traditionally measured on specialized ECG-gated cardiac CT, it can also be measured on nearly every standard CT scan of the chest performed without contrast[8]. However, the measurement requires radiological expertise, time, and specialized coronary calcium quantification equipment. As a result, this essentially free available information is usually not reported. An automated system for quantifying calcium on medical imaging could help put this actionable information into the hands of patients and their physicians.

Recent strides in artificial intelligence, deep learning in particular, have shown its viability in several medical applications such as medical diagnostic and imaging, risk management, or virtual assistants. Especially in medical imaging there is a large potential as deep learning can successfully be used for identifying and segmenting objects within the 3-dimensional image space[13–16]. A major advantage is that deep learning can automate complex assessments that previously could only be done by radiologists, but now is feasible at scale with a higher speed and lower cost. This makes deep learning a promising technology for automating cardiovascular event prediction from imaging. Before clinical introduction can be considered; however, generalizability of these systems needs to be demonstrated as they need to be able to predict cardiovascular events of asymptomatic and symptomatic individuals across multiple clinical scenarios, and work robustly on data from multiple institutions.

Here, we present a deep learning system that automatically and accurately can predict cardiovascular events by quantifying the presence and extent of coronary calcium. The system was evaluated in 20,084 individuals from four well-established prospective cohorts and randomized controlled trials - a healthy asymptomatic community-dwelling sample from the Framingham Heart Study (FHS)[17], older asymptomatic heavy smokers in the National Lung Screening Trial (NLST)[18], a symptomatic stable chest pain population evaluated for suspected coronary artery disease in the outpatient setting in the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE)[19], and a symptomatic acute chest pain population presenting to the emergency department in the Rule Out Myocardial Infarction using Computer Assisted Tomography (ROMICAT-II)[20] trial. Overall, the association between the algorithm's prediction and adverse cardiovascular events was tested in individuals who were imaged using different CT scanners, applying a variety of CT scan protocols, including ECG-gated and non-gated CT scans. Accuracy compared to the gold standard of expert human readers was assessed in 5,521 subjects across all four cohorts. Our results demonstrate that deep learning methods can automated cardiovascular risk predictions from medical images acquired in several clinical scenarios. These observations provide a rationale to implement this technique in both screening and hospital settings to improve population health, at high speed and low costs.

Results

We developed a deep learning system to automatically identify individuals at high risk for cardiovascular disease and tested the system's performance in four large independent held-out cohorts with a variety of clinical presentations and CT scanning techniques. Fig. 1 provides an overview of the test cohorts and analyses. Clinical characteristics of the test cohorts can be found in Table 1.

Table 1. Baseline characteristics of subjects in the four test cohorts.

Characteristics*	**FHS-CT2 (n=663)**	**NLST (n=14,959)**	**PROMISE (n=4,021)**	**ROMICAT-II (n=441)**
Woman - n (%)	372 (56.1)	6,110 (40.9)	2,047 (50.9)	235 (53.3)
Age - years	57.2±11.4	61.5±5.1	60.6±8.0	53.7±8.0
Body mass index - kg/m^2	28.6±5.5	27.9±5.1	30.4±5.9	29.3±5.2
Arterial hypertension - n (%)	219 (33.1)	5,321 (35.6)	2,614 (65.0)	233 (52.8)
Diabetes - n (%)	31 (4.87)	1,427 (9.5)	838 (20.8)	74 (16.8)
Hypercholesterolemia - n (%)	236 (35.6)	n/a	2,734 (68.0)	198 (44.9)
Former or current smoker - n (%)	202 (31.7)	14,959 (100)	2,078 (51.7)	220 (49.9)
Framingham risk score	0.10±0.1	n/a	0.22±0.2	n/a
TIMI risk score	n/a	n/a	n/a	0.13±0.3

FHS-CT2[17]: Framingham Heart Study, participants from the second examination cycle of the Third Generation Cohort; NLST[18]: National Lung Screening Trial; PROMISE[19]: Prospective Multicenter Imaging Study for Evaluation of Chest Pain; ROMICAT-II[20]: Rule Out Myocardial Infarction using Computer Assisted Tomography II; TIMI: Thrombolysis In Myocardial Infarction. n/a: Data was not available. *Characteristics are presented as mean ± standard deviation, if not stated otherwise.

Development of the deep learning system.

The Framingham Heart Study (FHS) is a long-term cardiovascular cohort study including asymptomatic persons originally from the city of Framingham in Massachusetts[17,21]. The Offspring and Third Generation FHS cohorts received ECG-gated non-contrast cardiac CT and were included in our analysis. We developed the deep learning system in the first cohort of FHS participants to have cardiac CT (FHS-CT1), including 1,636 individuals. The deep learning system was trained to identify and quantify coronary artery calcium based on manual segmentations performed by expert CT readers (Fig. 1a). To localize and segment the heart in a given CT scan, two consecutive deep learning networks were trained using 129 cardiac ECG-gated CTs with volumetric heart segmentations provided by expert readers. These networks were tested in an independent subset of our test cohorts including 1,857 cardiac gated and low-dose chest screening CT (Supplementary Fig. 2a). In this test cohort the heart localization network was able to predict the heart center with an accuracy of 9±7mm, while the heart segmentation network achieved a Dice coefficient of 0.90±0.059. Supplementary Table 4 provides details about the results of the two networks in the sub-cohorts. Next, the system automatically identified and segmented the coronary calcium and computed the coronary artery calcium (CAC) scores, and stratified them into clinically relevant categories of very low (CAC=0), low (CAC=1-100), moderate (CAC=101-300), and high (CAC>300). The system could analyze each image on an average of 1.938 seconds per scan on a GPU system. Resulting CAC scores were evaluated in the test cohorts in terms of agreement with expert readers as well as predicting risk of cardiovascular events on follow-up (Fig. 1b).

Automated coronary calcium scoring in lung cancer screening.

To evaluate the value of coronary calcium in heavy smokers having lung cancer screening CT, we applied our deep learning system to 14,959 participants in the low-dose chest CT arm of the National Lung Screening Trial (NLST). NLST low-dose chest CT was performed at 33 institutions with a variety of CT scanners using a non ECG-gated low dose chest CT protocol[18].

We investigated the association between our deep learning system's calcium score and incident atherosclerotic cardiovascular disease (ASCVD) death in lung cancer screening eligible individuals with a median follow-up time of 6.7 years. Kaplan-Meier analysis and Cox regression showed significant differences between all four calcium risk groups (Fig. 2a). Adjusted for age, sex, diabetes, heart disease, hypertension and stroke, the hazards ratio (HR) for cardiovascular disease compared to the reference (very low risk) group was 1.57 (95%CI=0.96-2.57, P=0.069) for the low risk group, 2.79 (95%CI=1.70-4.57, P<0.001) for the moderate risk group and 3.87 (95%CI=2.45-6.11, P<0.001) for the high risk group (Table 2).

Risk predictions in stable and acute chest pain patients.

Furthermore, we tested our deep learning system in outpatients with stable chest pain enrolled and randomized to ECG-gated cardiac CT in the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE). In 4,021 patients acquired at 193 North American sites, there was a graded association between extent of deep learning calcium score and cardiac events, defined as the composite of death, myocardial infarction, or hospitalization for unstable angina over median 25 months (P<0.001)(Fig. 2b)[19]. After adjustment for Framingham Risk Score, HRs for cardiac events showed significant increases in hazard across the low, moderate and high risk versus the reference (very low risk) group (Table 2).

The last test cohort included patients presenting with acute chest pain to the emergency department enrolled in the Rule Out Myocardial Infarction Using Computer Assisted Tomography II (ROMICAT-II) trial. In 441 patients who had ECG-gated cardiac CT at 9 sites, there was a similar association between the deep learning calcium score and acute coronary syndrome at 28 days (Fig. 2c)[20]. After adjustment for thrombolysis in myocardial infarction (TIMI) risk score, again patients with increasing deep learning calcium score were at increased risk, reflected in HRs significantly greater than 1 for each of low, moderate, high risk vs the reference (very low risk) group (Table 2). HRs increase as the risk category increases.

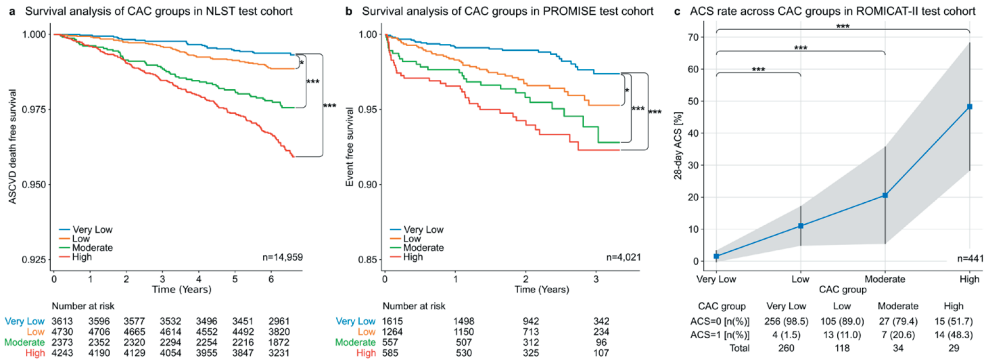


Figure 2. Outcome analysis for deep learning predicted calcium scores. Kaplan-Meier survival analysis of CAC risk groups for (**a**) cardiovascular disease-related death for 14,959 subjects of the National Lung Screening Trial (NLST)[18] and (**b**) all-cause mortality, myocardial infarction and unstable angina for 4,021 subjects of the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE)[19]. A two-sided log-rank test was used to calculate the p-values (*p-value ≤ 0.05; ***p-value ≤ 0.001) in panels (**a**) and (**b**). **c** Thirty-day acute coronary syndrome (ACS) rate across CAC groups for 441 subjects from the Rule Out Myocardial Infarction using Computer Assisted Tomography II (ROMICAT-II)[20] trial. The shaded area corresponds to the 95% confidence interval of the thirty-day ACS rate across CAC groups. A two-sided Fisher's exact test was used to estimate differences in the ACS rate between the very low risk group and the low, moderate and high risk group (***p-value ≤ 0.001). CAC: Coronary artery calcium; ASCVD: Atherosclerotic cardiovascular disease, ACS: Acute coronary syndrome. CAC risk groups: Very low: 0; Low: 1-100; Moderate: 101-300; High: >30021.

Comparison of automated deep learning and manual results.

We compared the deep learning calcium scores to manually measured calcium scores in 5,521 test cohort patients from FHS-CT2 (n=663), NLST (n=396), PROMISE (n=4,021) and ROMICAT-II (n=441). There was a very high[22] Spearman's correlation of 0.92 (P<0.0001) and substantial agreement[23] between automatically and manually calculated calcium risk groups (Fig. 3a). Most differences occurred between adjacent risk categories. For a detailed comparison of calcium scores in each test cohort, as well as concordance tables and kappa values, see Supplementary Figures 3, 4 and 5 and Supplementary Tables 1, and 3. Furthermore, an in-depth outlier-analysis was performed and can be found in the Supplementary Note 1.

To show the predictive value of the automatically calculated calcium score we computed the AUCs for event prediction in NLST, PROMISE and ROMICAT-II (Supplementary Table 2) and compared them to AUCs from manually derived calcium scores (Supplementary Table 5). We used random effects meta-analysis to estimate combined predicted and manual AUCs. The combined predicted AUC=0.74 was statistically not different to the combined manual AUC=0.75 (P=0.544).

Test-retest reliability.

A test-retest analysis was performed separately on the manual and on the deep learning risk scores on a subset of randomly selected 252 image pairs from FHS-CT1. Each image pair was taken consecutively within the same setup and within one minute to one-hour time difference. The results showed a great stability between the automatically calculated calcium scores for each image per pair achieving an ICC of 0.993 (P<0.001), compared to the ICC of manual calculated calcium scores of 0.997 (P<0.001). Manual and automatic test-retest repeatability is shown in Fig. 3b and 3c.

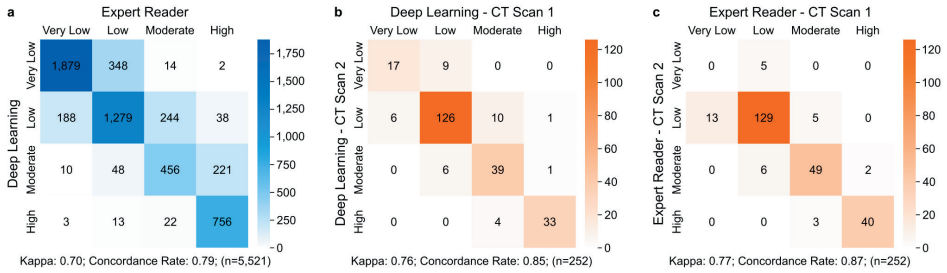


Figure 3. Confusion matrices to compare manual and automatic CAC quantification and to assess test-retest repeatability. a Comparison of CAC classes quantified by the deep-learning framework and expert readers combining data from FHS-CT2, NLST, PROMISE and ROMICAT-II (n=5,521). The robustness of **b** the deep learning framework and c expert readers to quantify CAC was assessed in 252 FHS-CT1 subjects who underwent two subsequent CT scans within one hour (Scan 1 and Scan 2). CAC: Coronary artery calcium; FHS-CT1[17], FHS-CT2[17]: Framingham Heart Study, (CT1) participants from the seventh examination cycle of the Offspring Cohort or first examination cycle of the Third Generation Cohort (2002-05) and (CT2) participants from the second examination cycle of the Third Generation Cohort (2008-11); NLST[18]: National Lung Screening Trial; PROMISE[19]: Prospective Multicenter Imaging Study for Evaluation of Chest Pain; ROMICAT-II[20]: Rule Out Myocardial Infarction using Computer Assisted Tomography II; CAC risk groups: Very low: 0; Low: 1-100; Moderate: 101-300; High: >300[21].

Discussion

In this investigation, we demonstrate that a deep learning based coronary calcium scoring system accurately stratifies the risk for cardiovascular events across 19,421 individuals with distinct presentations enrolled in four large clinical studies. Risk prediction was robust across multiple clinical scenarios, including a primary prevention asymptomatic setting with non-gated chest CT (NLST)[18], as well as dedicated ECG-gated cardiac CT in stable (PROMISE)[19] and acute (ROMICAT-II)[20] chest pain setting. The deep learning calcium score in 5,521 participants had high correlation with human expert readers and demonstrated robust test-retest reliability. Persons with a calcium score of zero are at very low risk[24], with increasing risk in the ordinal calcium score tiers identified by the deep learning system[25–27]. Based on the 2018 ACC/AHA guidelines[7], in persons at intermediate risk (defined as ≥7.5% to <19.9% 10-year risk of cardiovascular events based on risk factors), a calcium score of 0 indicates very low risk and unlikely benefit from statin therapy, while a high calcium score (≥100 or ≥ 75th centile for age/sex) indicates that a statin should be considered[7]. Despite these recommendations, dedicated coronary calcium scoring CT is not yet covered by Medicare and most US insurance companies, and for this reason, there is a great deal of interest in deriving the calcium score from routine chest CTs, which are far more common[8,28,29].

Traditionally, coronary calcium scoring requires special software, manual measurement by trained experts and dedicated ECG-gated cardiac CT. As a consequence, the calcium score is often not reported on routine noncardiac chest CT, despite the fact that calcium scores on non-gated CT have reasonably good agreement with dedicated calcium scoring CT[30,31]. Our automated calcium scoring system addresses this need by reliably and **Table 2.** Univariate and multivariable survival analyses of the predictive value of deep learning risk scores assessed in the test cohorts.

Table 2.

Risk Groups	Events*	Univariate			Multivariable					
		HR	95%CI	P-value	HR	95%CI	P-value	HR	95%CI	P-value
NLST: n=14,959; Events: ASCVD death, n=288										
					Adjusted for age and sex			Adjusted for age, sex, diabetes, hypertension, past heart-disease, past stroke		
Very low	0.6% (23/3613)	n/a	reference	reference	n/a	reference	reference	n/a	reference	reference
Low	1.1% (53/4730)	1.77	1.08-2.88	0.022	1.62	0.99-2.65	0.054	1.57	0.96-2.57	0.069
Moderate	2.4% (56/2373)	3.76	2.32-6.12	<0.001	3.05	1.87-5.00	<0.001	2.79	1.70-4.57	<0.001
High	3.7% (156/4243)	5.98	3.86-9.26	<0.001	4.34	2.75-6.84	<0.001	3.87	2.45-6.11	<0.001
PROMISE: n=4,021; Events: All-cause mortality, MI, UA, n=130										
					Adjusted for age and sex			Adjusted for Framingham Risk Score		
Very low	1.5% (25/1615)	n/a	reference	reference	n/a	reference	reference	n/a	reference	reference
Low	3.2% (41/1264)	2.16	1.31-3.55	0.002	1.96	1.18-3.25	0.009	1.90	1.15-3.15	0.012
Moderate	5.0% (28/557)	3.35	1.95-5.74	<0.001	2.83	1.62-4.96	<0.001	2.57	1.47-4.50	0.001
High	6.2% (36/585)	4.10	2.46-6.84	<0.001	3.21	1.84-5.60	<0.001	2.95	1.71-5.08	<0.001
ROMICAT-II: n=441; Events: ACS, n=38										
					Adjusted for age and sex			Adjusted for TIMI Risk Score		
Very low	1.5% (4/260)	n/a	reference	reference	n/a	reference	reference	n/a	reference	reference
Low	11.0% (13/118)	7.92	2.52-24.86	<0.001	6.46	1.96-21.30	0.002	7.70	2.44-24.30	0.001
Moderate	20.6% (7/34)	16.59	4.56-60.33	<0.001	12.60	3.19-49.83	<0.001	16.20	4.44-59.13	<0.001
High	48.3% (14/29)	59.73	17.50-203.78	<0.001	47.65	12.55-180.94	<0.001	57.11	16.55-197.12	<0.001

NLST[18]: National Lung Screening Trial; PROMISE[19]: Prospective Multicenter Imaging Study for Evaluation of Chest Pain; ROMICAT-II20: Rule Out Myocardial Infarction using Computer Assisted Tomography II; HR: Hazard ratio; CI: Confidence interval; CVD: Cardiovascular disease; ASCVD: Atherosclerotic cardiovascular disease; htn: hypertension; MI: Myocardial Infarction; UA: Unstable angina; ACS: Acute coronary syndrome; OR: Odds ratio; TIMI: Thrombolysis In Myocardial Infarction; n/a: Not available. Framingham Risk Score: Age, Total cholesterol, Smoker, HDL cholesterol systolic blood pressure. TIMI risk score: Age, Aspirin use, Angina, Elevated serum cardiac biomarkers, Known coronary artery disease, At least 3 risk factors for coronary artery disease, such as: Hypertension, Smoker, Low HDL cholesterol, Diabetes mellitus, Family history of premature coronary artery disease. Coronary calcium risk categories are based on: Very Low Risk (0: no coronary calcifications found), Low Risk (1-100: small amounts of coronary calcifications), Moderate Risk (101-300: moderate amounts of coronary calcifications), and High Risk (>300: large amounts of coronary calcifications)[21]. *Events are presented as a percentage within categorical risk group and with number of events and total number of subjects within the group in parenthesis.

accurately extracting the calcium score in both cardiac CT and chest CT. The system calculates the calcium score in under two seconds, without human input. Our approach has several innovations: first, we developed a unique deep learning system to measure coronary artery calcium on routine cardiac electrocardiography (ECG)-gated and non-gated chest CT, spanning a broad clinical spectrum including acute and stable chest pain as well as asymptomatic individuals having lung cancer screening. Our analysis of individuals from well-known NIH-sponsored observational cohorts and randomized controlled trials with prospective followup for cardiovascular events and death, is the largest to date to demonstrate the clinical value of automated calcium scoring. Second, we demonstrate prognostic value for risk of cardiovascular disease when the deep learning calcium score is applied to four different trials and longitudinal cohorts spanning the range of clinical scenarios in which coronary calcium would be useful. As our deep learning system does not require human input, this makes it an 'end-to-end' solution for accurate and time-efficient cardiovascular risk assessment in clinical settings[32,33]. Third, we share our rigorously validated deep learning system to the public, allowing for accelerated adoption of these technologies by both academic and commercial entities.

Although other studies have investigated deep learning algorithms for automated coronary calcium quantification[34–42], they used smaller cohorts or proprietary technologies. For example, previous publications for fully automatic coronary calcium assessment proposed models focused on either ECG-gated cardiac[34] or non-gated chest CT[35]. Shadmi et al.[40] trained slice based U-Net and FC-DenseNet networks to segment coronary calcium with high accuracy in a subset of NLST, optimizing their model for non-gated CT only. Lessmann at al. presented a two-stage approach for calcium scoring[37] as well as a deep learning method[35] in a smaller subset of NLST. Martin et al.[42] tested in their study a prototype commercial deep learning system for coronary calcium segmentation on a small data set from a single institution and scanner and their median computing time per scan was slightly slower (2.7 seconds). A combined solution capable of analyzing cardiac and non-gated chest CT as presented in our investigation has only been proposed by de Vos et al.[36] and van Velzen et al.[41]. The approach proposed by de Vos et al.[36] predicted the calcium score directly using direct regression on 2D CT slices only, and their test cohort was substantially smaller compared to our present analysis, less diverse and from the same sites as their training cohort[36]. Van Velzen et al.[41] has shown the automation of CAC measurements compared to manual assessments in several clinical scenarios, using a two step approach to find calcification candidates and subsequently detecting calcifications, again using smaller testing cohorts compared to our present study. Our analysis of 20,084 individuals from well-known observational cohorts and randomized controlled trials with prospective followup for cardiovascular events and death, is by far the largest to date to demonstrate the predictive value of automated calcium scoring. Furthermore we demonstrate strong robustness of the system by a high correlation with manual scoring in 5,521 subjects and high test-retest reliability in data from 252 individuals. We also share

our deep learning system, including the trained models, with the community, without restrictions.

To overcome different fields of view of CT scans in tested cohorts and to reduce the amount of data that has to be processed to assess coronary calcium, many approaches implement a preprocessing step to find a region of interest (ROI) around the heart. Often, traditional image processing techniques are used to find the ROI[38–40], but also 2D deep learning networks were successfully used to segment the heart and estimate a 'bounding box[43]. The benefit of our 3D heart segmentation step is not to find a rectangular ROI, but to narrow the region for coronary calcium segmentation to the heart itself.

A strength of our investigation was that we tested our system in populations from large clinical trials and longitudinal cohorts with well-adjudicated cardiovascular diseases events. This is essential, as before clinical introduction can be considered, generalizability of these automated systems needs to be demonstrated as they need to be able to predict cardiovascular events of asymptomatic individuals across multiple clinical scenarios and work robustly on data from multiple institutions. Overall, we included over 20,000 persons drawn from over 200 sites. The available health outcomes and risk factors varied between datasets, reflecting the diverse mix of asymptomatic and symptomatic individuals. Nevertheless, the deep learning calcium score was an independent predictor of adverse cardiovascular events in all cohorts. The majority of FHS (100%)[44], NLST (91%)[18], PROMISE (77%)[45] and ROMICAT-II (66%)[46] participants were non-Hispanic whites. Although the manual calcium score has proven to be an important predictor of cardiovascular events across race and ethnicities, generalizability to other demographics will need to be investigated in future studies[47]. Furthermore, the proposed system evaluated the coronary artery calcium score on non-contrast cardiac and chest CT. As such it could not detect noncalcified plaque, which can be present even with a calcium score of zero.

In summary, our end-to-end deep learning system provides an automated quantification of coronary calcium on both cardiac CT and lung cancer screening CT. The deep learning calcium score is strongly associated with cardiovascular risk in a broad spectrum of clinical scenarios. Automated quantification of coronary calcium has the potential to improve clinical routine and population health.

Methods

Study population

This study was a retrospective secondary analysis of a longitudinal primary prevention cohort (FHS-CT1 and FHS-CT2) and three randomized clinical trials (NLST, PROMISE, ROMICAT II). Details about participant selection are provided in the consort diagrams in the Supplementary Fig. 1.

The deep learning system training and tuning was accomplished in Framingham Heart Study (FHS) Offspring[48] and Third Generation[49] cohort participants (FHS-CT1,

n=1,636) who had non-contrast ECG-gated cardiac CT for coronary calcium quantification between 2002 and 2005. Details regarding the FHS cohort, inclusion criteria, and calcium scoring have been described elsewhere[21]. Participants resided or had parents who resided in Framingham or in the New England region. Major inclusion criteria were age ≥35 years for men and ≥40 years for women. All participants provided written consent for the CT study, which was approved by the institutional review boards of the Boston University Medical Center and Massachusetts General Hospital[17,21]. In our investigation we included only participants with available cardiovascular disease risk profile, no known prior cardiovascular disease, and diagnostic-quality cardiac CT as determined by an expert reader (Supplementary Fig. 1a).

The deep learning system performance was tested in a second, independent group of FHS participants who had cardiac CT from 2008-2011 (FHS-CT2, n=663). None of the persons in the FHS-CT2 testing cohort were in the FHS-CT1 training/tuning cohort. While the FHS-CT1 training cohort included only diagnostic-quality CTs, the FHS-CT2 testing cohort included all CTs including those initially considered non-diagnostic (Supplementary Fig. 1b).

A second testing cohort was drawn from the National Lung Screening Trial (NLST)[18], a multicenter randomized controlled trial of non-contrast, non ECG-gated low-dose chest CT for lung cancer screening. In NLST, 53,454 subjects aged 55-74 years, current or former heavy smokers, were enrolled at 33 participating medical institutions with all-cause mortality as primary outcome measure over a follow up of up to 8 years. 26,722 randomly selected participants underwent low-dose non-contrast chest CT imaging between 2002 and 2007. The trial was approved by the institutional review board at each site. From the full cohort we had permission to include 15,000 randomly selected subjects. For each subject the baseline (T0) CT scan was chosen with soft kernel preferred over hard kernel reconstructed images. We excluded 17 subjects that did not have a T0 scan, 12 subjects that did not have scans which met our quality requirements, 12 subjects that had a broken or incomplete scan and 17 subjects with missing risk data. The final testing cohort consisted of 14,959 scans (Supplementary Fig. 1c). To verify the results of our deep learning system in this cohort, a subset of randomly chosen 396 subjects were segmented by expert readers.

The third cohort included participants from the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE)[19,45]. In this multicenter trial 10,003 symptomatic patients were randomized at 193 medical sites in North America using a composite of major cardiovascular events as a primary outcome measure over a median follow up of 25 months. Participants of age 45 to 64 years with stable chest pain and without known prior CAD were enrolled between 2010 and 2013 with 4,996 subjects randomly selected to undergo cardiac CT imaging. The central activities of the study were approved by the Duke, Partners Healthcare and Tufts Institutional Review Boards. Furthermore, local or central IRBS approved the study at each medical institution. The final

testing cohort included 4,021 individuals each with a non-contrast cardiac CT scan and full risk profile available (Supplementary Fig. 1d). All subjects were segmented by expert readers.

The fourth cohort included participants from the Rule Out Myocardial Infarction Using Computer Assisted Tomography Study Two (ROMICAT-II)[20]. In this randomized open-labeled multicenter trial 1000 patients which presented at the emergency department of nine clinical sites with acute chest pain were enrolled between 2010 and 2012. The primary outcome measure of this study was the length of the hospital stay and a second outcome including undetected acute coronary syndrome within 72 hours after hospital discharge, increased adverse events, major adverse cardiovascular events within 28 days, and periprocedural complications. The study was approved by the local institutional review boards. Patients were between 40 and 74 years old, without known coronary artery disease, almost equal gender representation and significant representation of all minorities. Of these subjects, 500 were randomly selected to undergo non-contrast cardiac CT imaging. After excluding participants with incomplete image data or risk profile the final testing cohort included 441 participants (Supplementary Fig. 1e). All subjects were segmented by expert readers.

A detailed population description for all four cohorts can be found in Table 1. Participants from all studies provided written consent.

Deep learning based coronary calcium segmentation

We propose a deep learning system which is able to automatically calculate a calcium score from a given CT scan for cardiovascular risk prediction. The system consists of four consecutive steps for (1) heart localization, (2) heart segmentation, (3) coronary calcium segmentation, and (4) calcium score calculation. We trained a separate fully convolutional neural network of the U-Net[50] architecture for each of the first three steps. The U-Net architecture was originally designed for biomedical image segmentation with the goal of overcoming the requirement for a very large cohort for training a deep learning network.

The cohort for training and tuning the three deep learning models consisted of 1,636 CT scans: 623 CT scans were from subjects with coronary calcium and 1,013 CT scans of subjects with no coronary calcium. Although several hundred more CT scans from subjects with no coronary calcium were available, we chose to exclude them to keep the imbalance between subjects with and without coronary calcium small. Excluded subjects were selected randomly. Coronary calcium, if present, was manually segmented by experienced readers in all subjects. Furthermore, the heart was manually segmented in a subset of 129 randomly selected subjects of the training cohort. Our testing cohort consisted of 20,084 subjects from four different clinical studies and trials including dedicated cardiac CT scans as well as lung screening CT scans, health outcomes, and follow up information. Manually calculated calcium scores from expert readers were available for 5,521 subjects and manually segmented hearts for 895 subjects. All CT scans

were padded and cropped to the same size of 512ˣ512ˣ512pixel (px) and resampled to the same resolution of 0.7ˣ0.7ˣ2.5mm/px. A detailed description of the training, tuning and testing cohorts and their usage is described in the Supplementary Fig. 2a.

The first network in our system was trained to localize the heart within a given 3-dimensional (3D) CT scan. This step was necessary as CT scans can differ, for example, in size, resolution, area captured, or field of view, depending on the cohort, scanner used, and site acquiring the scan. The training cohort was split 70/30% for training and tuning, and all scans were down-sampled to a size of 112ˣ112ˣ112px to fit into the GPU (Graphics Processing Unit) memory. The model used for training was a standard U-Net with four down-sampling steps running for 1200 epochs. Data augmentation was used by applying rotation of ±4 degrees around the sagittal, transversal and longitudinal axis for heart localization and ±35 degrees around the sagittal axis for heart segmentation. Furthermore, we applied translation within ±10px in the axial plane for heart localization and ±20px in the axial plane for heart segmentation. The output of the network was up-sampled to the initial CT scan size leading to a very rough heart segmentation which we used for placing a bounding box for the subsequent steps.

The second network of the deep learning system was trained to segment the heart. The input scans were first cropped to 384ˣ384ˣ80px cubes around the heart center and then down-sampled to 128ˣ128ˣ80px. The training cohort was again split 70/30% for training and tuning and data was augmented by applying rotations and translations in small ranges only. The model used for training had the same architecture as in the previous step with four down-sampling steps running for 1000 epochs. Once the model parameters were found to perform well on the tuning cohort, the final model was trained combining the training and tuning cohort for better performance. The output of the network was up-sampled to initial CT scan size leading to an accurate heart segmentation. As this step was mainly to reduce the area for the consecutive calcium segmentation step and although the error of the heart segmentation was low, we added a rim of 11 pixel to the predicted heart segmentation to ensure the whole heart was captured.

The third network was trained to segment coronary calcium. For this step, we divided the previously segmented heart into smaller cubes of size 48ˣ48ˣ32px. Extensive testing of several cube sizes showed the chosen size worked best as larger cubes increased training time while the accuracy stayed the same. We used cubes overlapping all but one pixel for the training, whereas cubes did not overlap during the testing. The model used in this step was a U-Net[50] with three down-sampling steps extended by batch normalization layers in the contracting path (left side) for better generalizability (Supplementary Fig. 2b). The resulting segmentation patches were aligned again leading to a coronary calcium segmentation of the heart. The final step was to threshold the whole segmentation by 0.95 to obtain the binary calcium mask.

With the coronary calcium segmented the calcium score was calculated using a volumetric implementation of the method by Agatston and Janowitz[5]. The calcium

score was calculated by multiplying the volume of a coronary calcification with a factor, depending on the highest density within the calcification, with the density being measured in Hounsfield Units (HU). This weight factor was 1 for a density between 130 and 199 HU, 2 for 200 to 299 HU, 3 for 300 to 399 HU, and 4 for 400 HU and above. Calcification with a volume below one cubic-millimeter was considered noise and excluded from the calculation. The final calcium score per patient was the sum of the weighted calcifications. For further analysis we stratified the calcium risk score into the risk groups very low (0), low (1-100), moderate (101-300) and high (>300)[21].

Training, tuning and testing was done on a Linux workstation using Tensorflow-GPU and Keras. The only notable hardware requirement was to have at least 64 gigabyte of GPU memory to fit a reasonable batch-size of input volumes for the heart segmentations.

Technical Evaluation

The performance of the deep learning system was tested by reviewing CT scans with discordances between the manual and deep learning coronary calcium categories (Supplementary Fig. 7). Most discrepancies were due to misclassification of non-coronary calcium as coronary calcium and vice versa. In a few instances, inaccurate heart segmentation led to coronary calcium being outside the region of interest of the calcium segmentation network and hence being missed. Furthermore, we measured the time the system needed to process scans. On average, the deep learning system assessed the coronary artery calcium score in under two seconds per scan.

Manual calcium score assessment

The coronary artery calcium score was measured manually by expert readers using the method by Agatston and Janowitz[5,6] on dedicated workstations, as reported in the parent FHS[17], PROMISE[19] and ROMICAT-II[20] studies. In NLST[18], the coronary calcium score was measured manually in 396 randomly selected participants using 3D Slicer (V4).

Test-Retest analysis

In FHS-CT1, 252 participants underwent cardiac ECG-gated CT twice within one hour, to assess test retest reliability. The deep learning system and the human readers quantified calcium on both scans to assess test-retest reliability.

Statistical methods

In this study we described continuous variables as the mean ± standard deviation (SD) and categorical variables as frequencies and percentages. Furthermore, we performed univariate and multivariate Cox proportional hazards regressions comparing cardiovascular disease risk in the 1st (lowest) vs. the 2nd–4th quartiles of CAC. The dependent variable in FHS-CT2 was a composite of cardiovascular disease event and all-cause mortality with a mean follow-up of 8.8 years[51]. Events in NLST were defined

as atherosclerotic cardiovascular disease (ASCVD) mortality with a follow-up of up to 9 years[18]. Events in PROMISE were defined as a composite of all-cause mortality, myocardial infarction, major complications from cardiovascular procedures and diagnostic testing and unstable angina with an average follow-up of 2.5 years[19]. Events in the ROMICAT-II trial were defined as major adverse cardiovascular events in a time frame of 28 days after admission to the emergency room[46]. Cox proportional hazards models and log rank tests were used to estimate and compare hazard ratios between the reference (very low risk) group (0: no coronary calcifications found), the low risk group (1-100: small amounts of coronary calcifications), the moderate risk group (101-300: moderate amounts of coronary calcifications) and the high risk group (>300: large amounts of coronary calcifications): one unadjusted model, one model adjusted for age and sex, and a third model in which we additionally adjusted for standard cardiovascular risk factors (NLST: hypertension, diabetes, past coronary artery disease, past stroke; PROMISE: Framingham risk score (FRS); ROMICAT-II: Thrombolysis In Myocardial Infarction (TIMI) risk score) using Survival R package (v3.2-3). Standard Kaplan-Meier survival curves were generated to visualize event-free survival for the NLST and PROMISE testing cohorts in R using the Survminer package (v0.4.8). The log rank test was used to identify significant differences in survival. To assess the similarity of automatically and manually derived calcium scores, we calculated the Spearman's correlation (Python package scipy.stats.spearmanr based on Zwillinger, D. and Kokoska[52]), the intra-class correlation (ICC) which was calculated from components of a one-way analysis of variance in R using the ICC package (v2.3.0) based on Searle[53], Donner[54] and Thomas and Hultquist[55], the Cohen's Kappa[56] (Python package sklearn.metrics.cohen_kappa_score) and the concordance rate (calculated as number of concordant pairs divided by the number of all pairs). The combined AUCs were estimated and compared using the survcomp R package (V1.36.1). BMI was calculated using the weight (pounds) and height (inches) as: Weight / Height^2 * 703.

Acknowledgements

General: The authors thank the Framingham Heart Study, NCI, ACRIN, NLST, Prospective Multicenter Imaging Study for Evaluation of Chest Pain, and Rule Out Myocardial Infarction Using Computer Assisted Tomography II trial for access to trial data.

Funding

The authors acknowledge financial support from NIH (HA: NIH-USA U24CA194354, NIH-USA U01CA190234, NIH-USA U01CA209414, and NIH-USA R35CA22052; UH: NIH, 5R01-HL109711, NIH/NHLBI 5K24HL113128, NIH/NHLBI 5T32HL076136, NIH/NHLBI 5U01HL123339), the European Union - European Research Council (HA: 866504), as well as the German Research Foundation (DFG; TA: 1438/1-1 and WE: 6405/2-1), American Heart Association Institute for Precision Cardiovascular Medicine (MTL: 18UNPG34030172),

Fulbright Visiting Researcher Grant (E0583118), Rosztoczy Foundation Grant. The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the National Heart, Lung and Blood Institute.

Author contributions

Study design: R.Z., B.F., .P.E., J.W., I.A., J.T., C.H., U.H., M.T.L., H.J.W.L.A.; Code design, implementation and execution: R.Z.; Acquisition, analysis or interpretation of data: R.Z., B.F., .P.E.; Image segmentation: B.F., P.E., J.W., I.A., J.T., R.M.A., D.B., M.U., Y.K., J.K., L.Z., J.E.S.; Writing of the manuscript: R.Z., B.F., J.W., U.H., M.T.L., H.J.W.L.A.; Critical revision of the manuscript for important intellectual content: All authors; Statistical Analysis: R.Z., B.F., M.T.L., A.L., J.M.M., T.M.; Study supervision: U.H., M.T.L., H.J.W.L.A.

Competing interests

The authors declare no competing interests to this work.

Data availability

NLST data including raw CT images may be requested from the National Cancer Institute (https://biometry.nci.nih.gov/cdas/nlst/). Although raw CT imaging data cannot be shared, all measured results to replicate the statistical analysis are shared at the AIM webpage at aim.hms.harvard.edu/deepcac. Furthermore, we include test samples from a publicly available data set with deep learning and expert reader heart and calcium segmentations.

Code availability

The code of the deep learning system, as well as the trained model and statistical analysis are publicly available at the AIM webpage aim.hms.harvard.edu/deepcac.

References

1. Wilkins, E. *et al. European Cardiovascular Disease Statistics 2017*. (2017).
2. Writing Group Members *et al.* Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. *Circulation* **133**, e38–360 (2016).
3. Pokharel, Y. *et al.* Adoption of the 2013 American College of Cardiology/American Heart Association Cholesterol Management Guideline in Cardiology Practices Nationwide. *JAMA Cardiol* **2**, 361–369 (2017).
4. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* vol. 2 158–164 (2018).
5. Agatston, A. S. *et al.* Quantification of coronary artery calcium using ultrafast computed tomography. *J. Am. Coll. Cardiol.* **15**, 827–832 (1990).
6. Thanassoulis, G. *et al.* A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. *Circ. Cardiovasc. Genet.* **5**, 113–121 (2012).
7. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. *Journal of the American College of Cardiology* vol. 73 3168–3209 (2019).
8. Hecht, H. S. *et al.* 2016 SCCT/STR guidelines for coronary artery calcium scoring of noncontrast noncardiac chest CT scans: A report of the Society of Cardiovascular Computed Tomography and Society of Thoracic Radiology. *J. Thorac. Imaging* **32**, W54–W66 (2017).
9. Patel, M. R. *et al.* Correction to: ACC/AATS/AHA/ASE/ASNC/SCAI/SCCT/STS 2017 Appropriate Use Criteria for Coronary Revascularization in Patients With Stable Ischemic Heart Disease. *J. Nucl. Cardiol.* **25**, 2191–2192 (2018).
10. Raff, G. L. *et al.* SCCT guidelines on the use of coronary computed tomographic angiography for patients presenting with acute chest pain to the emergency department: A Report of the Society of Cardiovascular Computed Tomography Guidelines Committee. *Journal of Cardiovascular Computed Tomography* vol. 8 254–271 (2014).
11. Ravenel, J. G. & Nance, J. W. Coronary artery calcification in lung cancer screening. *Translational Lung Cancer Research* vol. 7 361–367 (2018).
12. Gupta, A. *et al.* The Identification of Calcified Coronary Plaque Is Associated With Initiation and Continuation of Pharmacological and Lifestyle Preventive Therapies: A Systematic Review and Meta-Analysis. *JACC Cardiovasc. Imaging* **10**, 833–842 (2017).
13. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
14. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).

15. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
16. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
17. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
18. National Lung Screening Trial Research Team *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011).
19. Douglas, P. S. *et al.* PROspective Multicenter Imaging Study for Evaluation of chest pain: rationale and design of the PROMISE trial. *Am. Heart J.* **167**, 796–803.e1 (2014).
20. Hoffmann, U. *et al.* Design of the Rule Out Myocardial Ischemia/Infarction Using Computer Assisted Tomography: a multicenter randomized comparative effectiveness trial of cardiac computed tomography versus alternative triage strategies in patients with acute chest pain in the emergency department. *Am. Heart J.* **163**, 330–8, 338.e1 (2012).
21. Hoffmann, U., Massaro, J. M., Fox, C. S., Manders, E. & O'Donnell, C. J. Defining normal distributions of coronary artery calcium in women and men (from the Framingham Heart Study). *Am. J. Cardiol.* **102**, 1136–41, 1141.e1 (2008).
22. Mukaka, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **24**, 69–71 (2012).
23. Viera, A. J. & Garrett, J. M. Understanding interobserver agreement: the kappa statistic. *Fam. Med.* **37**, 360–363 (2005).
24. Blaha, M. J. *et al.* Role of Coronary Artery Calcium Score of Zero and Other Negative Risk Markers for Cardiovascular Disease: The Multi-Ethnic Study of Atherosclerosis (MESA). *Circulation* **133**, 849–858 (2016).
25. Hoffmann, U. *et al.* Cardiovascular Event Prediction and Risk Reclassification by Coronary, Aortic, and Valvular Calcification in the Framingham Heart Study. *J. Am. Heart Assoc.* **5**, e003144 (2016).
26. Budoff, M. J. *et al.* Ten-year association of coronary artery calcium with atherosclerotic cardiovascular disease (ASCVD) events: the multi-ethnic study of atherosclerosis (MESA). *Eur. Heart J.* **39**, 2401–2408 (2018).
27. Mitchell, J. D., Paisley, R., Moon, P., Novak, E. & Villines, T. C. Coronary Artery Calcium and Long-Term Risk of Death, Myocardial Infarction, and Stroke: The Walter Reed Cohort Study. *JACC Cardiovasc. Imaging* **11**, 1799–1806 (2018).
28. Hecht, H. S. Coronary Artery Calcium Analysis and Reporting on Noncontrast Chest CT Scans: a Paradigm Shift in Prevention. *Current Cardiovascular Imaging Reports* vol. 9 (2016).

29. Lu, M. T. *et al.* Lung Cancer Screening Eligibility in the Community: Cardiovascular Risk Factors, Coronary Artery Calcification, and Cardiovascular Events. *Circulation* **134**, 897–899 (2016).
30. Huang, Y.-L. *et al.* Reliable categorisation of visual scoring of coronary artery calcification on low-dose CT for lung cancer screening: validation with the standard Agatston score. *Eur. Radiol.* **23**, 1226–1233 (2013).
31. Budoff, M. J. *et al.* Coronary artery and thoracic calcium on noncontrast thoracic CT scans: comparison of ungated and gated examinations in patients from the COPD Gene cohort. *J. Cardiovasc. Comput. Tomogr.* **5**, 113–118 (2011).
32. Takx, R. A. P. *et al.* Quantification of coronary artery calcium in nongated CT to predict cardiovascular events in male lung cancer screening participants: Results of the NELSON study. *Journal of Cardiovascular Computed Tomography* vol. 9 50–57 (2015).
33. Chiles, C. *et al.* Association of Coronary Artery Calcification and Mortality in the National Lung Screening Trial: A Comparison of Three Scoring Methods. *Radiology* **276**, 82 (2015).
34. Wolterink, J. M. *et al.* Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med. Image Anal.* **34**, 123–136 (2016).
35. Lessmann, N. *et al.* Sex Differences in Coronary Artery and Thoracic Aorta Calcification and Their Association With Cardiovascular Mortality in Heavy Smokers. *JACC Cardiovasc. Imaging* (2019) doi:10.1016/j.jcmg.2018.10.026.
36. de Vos, B. D. *et al.* Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT. *IEEE Trans. Med. Imaging* (2019) doi:10.1109/TMI.2019.2899534.
37. Lessmann, N. *et al.* Automatic Calcium Scoring in Low-Dose Chest CT Using Deep Neural Networks With Dilated Convolutions. *IEEE Trans. Med. Imaging* **37**, 615–625 (2018).
38. Huo, Y. *et al.* Coronary calcium detection using 3D attention identical dual deep network based on weakly supervised learning. *Medical Imaging 2019: Image Processing* (2019) doi:10.1117/12.2512541.
39. Santini, G. *et al.* An automatic deep learning approach for coronary artery calcium segmentation. *EMBEC & NBC 2017* 374–377 (2018) doi:10.1007/978-981-10-5122-7_94.
40. 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)* (2018) doi:10.1109/isbi.2018.8363515.
41. van Velzen, S. G. M. *et al.* Deep Learning for Automatic Calcium Scoring in CT: Validation Using Multiple Cardiac CT and Chest CT Protocols. *Radiology* **295**, 66–79 (2020).

42. Martin, S. S. *et al.* Evaluation of a Deep Learning–Based Automated CT Coronary Artery Calcium Scoring Algorithm. *JACC: Cardiovascular Imaging* vol. 13 524–526 (2020).
43. Lessmann, N. *et al.* Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest CT. *Medical Imaging 2016: Computer-Aided Diagnosis* (2016) doi:10.1117/12.2216978.
44. Tsao, C. W. & Vasan, R. S. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *Int. J. Epidemiol.* **44**, 1800–1813 (2015).
45. Douglas, P. S. *et al.* Outcomes of anatomical versus functional testing for coronary artery disease. *N. Engl. J. Med.* **372**, 1291–1300 (2015).
46. Hoffmann, U. *et al.* Coronary CT angiography versus standard evaluation in acute chest pain. *N. Engl. J. Med.* **367**, 299–308 (2012).
47. Detrano, R. *et al.* Coronary Calcium as a Predictor of Coronary Events in Four Racial or Ethnic Groups. *New England Journal of Medicine* vol. 358 1336–1345 (2008).
48. Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. An investigation of coronary heart disease in families. The Framingham offspring study. *Am. J. Epidemiol.* **110**, 281–290 (1979).
49. Splansky, G. L. *et al.* The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, Recruitment, and Initial Examination. *Am. J. Epidemiol.* **165**, 1328–1335 (2007).
50. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Lecture Notes in Computer Science* 234–241 (2015).
51. Qazi, S. *et al.* Increased Aortic Diameters on Multidetector Computed Tomographic Scan Are Independent Predictors of Incident Adverse Cardiovascular Events: The Framingham Heart Study. *Circ. Cardiovasc. Imaging* **10**, (2017).
52. Kokoska, S. & Zwillinger, D. *CRC Standard Probability and Statistics Tables and Formulae, Student Edition*. (CRC Press, 2000).
53. Ahrens, H. Searle, S. R.: Linear Models. John Wiley & Sons, Inc., New York-London-Sydney-Toronto 1971. XXI, 532 S. $9.50. *Biometrische Zeitschrift* vol. 16 78–79 (1974).
54. Donner, A. THE USE OF CORRELATION AND REGRESSION IN THE ANALYSIS OF FAMILY RESEMBLANCE. *American Journal of Epidemiology* vol. 110 335–342 (1979).
55. Thomas, J. D. & Hultquist, R. A. Interval Estimation for the Unbalanced Case of the One-Way Random Effects Model. *The Annals of Statistics* vol. 6 582–587 (1978).
56. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* vol. 20 37–46 (1960).

Chapter 3

Deep Learning for fully automatic heart segmentation in computed tomography scans

Roman Zeleznik, Borek Foldyna, Jakob Weiss, Parastou Eslami, Dennis Bontempi, Pamela S. Douglas, Ramachandran S. Vasan, Udo Hoffmann, Michael T. Lu, Hugo J.W.L. Aerts

Submitted (2021)

Abstract

Deep learning is able to automate complex tasks such as whole heart segmentation in computed tomography (CT), which is time consuming and requires dedicated expertise and equipment. Here we present an open-source deep learning system for fully automatic heart segmentation in cardiac ECG-gated and non-gated CT. The system was developed using 2,162 cardiac CT scans from the Framingham Heart Study (n=628) and the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE, n=1,140), as well as low-dose chest screening CTs from the National Lung Screening Trial (NLST, n=394), with manual segmentations provided by experienced medical readers. The deep learning system was able to accurately and precisely localize and segment the heart (Median Dice: 0.95, IQR: 0.02; Median surface distance: 1.6mm, IQR:0.71) in 1.2 seconds per scan. The mean Spearman's Correlation between the volumes of automatic and manual heart segmentations was very high (0.96, p<0.0001). We demonstrate the generalizability of a deep learning system for whole heart segmentation in large and distinctive test cohorts. Providing such an accurate and reliable open source system to the public has the potential to accelerate experimental studies and potential clinical application.

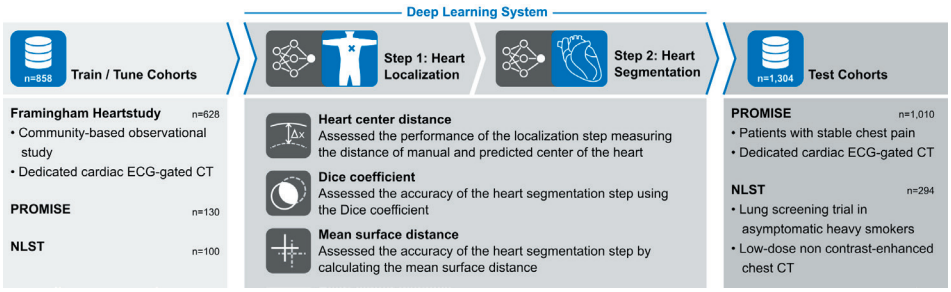


Figure 1. Overview of the proposed deep learning system for fully automatic heart segmentation in Computed Tomography (CT). The system was trained on cardiac ECG-gated CT scans from the Framingham Heartstudy as well as small subsets of the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) and low-dose chest screening CT from the National Lung Screening Trial (NLST) with manual heart segmentations provided by expert readers. It comprises a two step approach for localizing and subsequently segmenting the heart. The test cohort included cardiac-ECG gated CT from the PROMISE and low-dose chest screening CT from NLST.

Introduction

Deep learning has made substantial improvements in recent years, utilizing the high computational power of state-of-the-art computer hardware and the availability of massive amounts of digital data to match and even surpass human performance in task-specific applications (Mnih et al., 2015; Moravčík et al., 2017; Silver et al., 2017; Xiong et al., 2017). For instance, deep learning has shown promising results in the field of computer graphics and consequently in medical imaging, including radiology and dermatology (Esteva et al., 2017; Hosny et al., 2018). More specifically, deep learning has been successfully applied to image segmentation tasks where U-Nets and U-Net-derived networks are prominent examples for medical image segmentation tasks such as brain tumor segmentation (Dong et al., 2017) or lung nodule segmentation (Tong et al., 2018). These advances have led to a vast amount of proof-of-principle studies in various clinical application areas (Mahbod et al., 2018; Um et al., 2020; Zhuang et al., 2019). However, generalizability of such systems were not often demonstrated due to a lack of enough data for development and/or independent validation.

One important application of deep learning is fully automatic whole heart segmentation in computed tomography (CT), which is a critical task performed in medical research and care (Dimopoulos et al., 2013; Mohan et al., 1988). Moreover, heart segmentation represents the foundation for a wide range of other applications, such as coronary calcium segmentation (Zeleznik et al., 2021a) or quantification of pericardial fat (Goeller et al., 2018). A publicly available, robust and fully automatic heart segmentation system has the potential to accelerate scientific research and may allow for translation to medical care.

Here we propose a reliable and accurate deep learning system that is able to automatically segment the heart in non-contrast enhanced cardiac ECG-gated CT and low-dose chest screening CT scans. The system was developed with 2,162 CT scans, including scans from the Framingham Heart Study (FHS) (D'Agostino et al., 2008; Hoffmann et al., 2008), the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) (Douglas et al., 2014), and the National Lung Screening Trial (NLST) (National Lung Screening Trial Research Team et al., 2011), with manual segmentations provided by experienced medical readers.

Our test cohorts included 1,304 CT scans acquired at 226 medical sites. We assessed performance, generalizability, and applicability of the proposed deep learning system for heart segmentation in medical images, making it suitable for a vast array of research and medical applications. By making the code open-source and providing the trained model to the public without restrictions, our investigation has the potential to accelerate clinical research and medical treatment

Materials and Methods

Data

The training and tuning cohort included high quality cardiac ECG-gated CT from the Framingham Heart Study (FHS, n=628). To increase the training sample size, we also included 130 cardiac CTs from PROMISE as well as 100 low-dose chest screening CTs from NLST. The testing cohort included 1,010 cardiac CTs from PROMISE as well as non-gated CTs from NLST, acquired in 226 different medical sites (PROMISE: n=193, NLST: n=33). Manual heart segmentations for all cohorts were provided by the Cardiovascular Imaging Research Center (CIRC) at the Massachusetts General Hospital and Harvard Medical School in Boston. Participants from all studies provided written informed consent, which was approved by the institutional review boards of the Boston University Medical Center, Massachusetts General Hospital and Harvard Medical School. Further details regarding the cohorts and inclusion criteria have been described elsewhere (D'Agostino et al., 2008; Douglas et al., 2014; Hoffmann et al., 2008; National Lung Screening Trial Research Team et al., 2011). Additional information about the data selection criteria can be found in the Supplementary Methods 1.

Manual heart segmentations

Several independent cardiovascular experts with experience in cardiac imaging segmented the heart in the CT scans using a dedicated workstation (3DSlicer, v.4.7.0) ("3D Slicer," n.d.). Hearts were segmented as a 3D volume of the pericardial sac, including all four chambers (i.e., ventricles and atria), walls, coronary arteries, and epicardial fat. The cross-section of the right mid pulmonary artery represented the cranial border of the heart volume, while the left-ventricular apex represented the most caudal axial slice. Anterior, posterior, and lateral borders of the cardiac volume were defined by the pericardium. This method represents a standard as described in previous studies investigating heart volume and epicardial fat (Lu et al., 2016). To segment the heart volume, the readers manually traced the pericardium in axial slices in 10mm intervals and interpolated all missing slices. Individual borders of the interpolated slices were reviewed by the readers and adjusted if necessary.

Deep Learning System

We propose a deep learning system that consists of two consecutive steps for fully automatic heart segmentation in non-contrast enhanced CTs. The first step is a preparation step, used to localize the heart and crop the images to the area around the heart. Such processing helps to accommodate for the variability of scans in terms of scan size, resolution, field of view, and overall image quality and furthermore reduces the amount of data needed to be processed in the following step. The second step segments the heart in the cropped scans. For each of the two steps we used a customized 3-dimensional (3D)

implementation of the U-Net DL model (Ronneberger et al., 2015). During training and tuning, both networks were provided with the manual segmentations, while for testing the output of the first network was used as input for the second network. An overview of the full pipeline is shown in the Supplementary Figure S1.

Image preprocessing

With CT scans being acquired at different medical sites using varying CT protocols, the scan's field of view and size had to be normalized. Therefore, all scans were resampled to the most common resolution in the cohorts to an axial in-plane resolution of 0.68mm per pixel (pixel-spacing) and 2.5mm between each image plane (slice-thickness). Afterwards, the scans were re-sized to the same in-plane size of 512x512pixel while keeping the number of slices at this point unchanged. To achieve this, bigger scans were cropped to the center while smaller scans were padded with the smallest value defined for DICOM files (-1,024 HU). The scans were converted from the DICOM to the NRRD file format for easier file handling.

Deep learning - Heart localization (Step 1)

To localize the heart in a full sized body CT scan, we downsampled the scans and manual segmentation masks to avoid running into GPU memory limitations while training the deep learning model. As the heart is still visible even in images of very low resolution, it was possible to reliably generate a coarse segmentation mask in the downsampled scans. Although the generated masks were too rough to be used as a final result, once upsampled to the original size, they could be used to reliably compute the center of the heart.

The network we used in this step was a 3D U-Net with 4 downsampling steps (Ronneberger et al., 2015). Furthermore, we implemented a feature reduction step in the bottleneck layer to further reduce the memory allocation of the model. This also allowed us to reduce training time without decreasing the network's performance. The input scans and segmentation masks were down-sampled to 3.0mm in all directions, cropped to cubes with 112pixel edge length, and afterwards split randomly into a training set (70%) and tuning set (30%). To artificially increase the training data we augmented the scans by applying translations within ±10px in the axial plane for heart localization and rotations of ±4 degrees around the sagittal, transversal and longitudinal axis. The output of the network was a probability mask which was thresholded by a factor of 0.99 to a binary segmentation mask and subsequently upsampled to the original scan size. The model was trained for 1,200 epochs with an initial learning rate of 10-5 and a fixed learning rate drop of 50% every 100 epochs.

Deep learning - Heart segmentation (Step 2)

A second deep learning network was used for the final high resolution heart segmentation, which required the input data to be cropped to the heart. The minimum bounding-box size was calculated using the size of the manual segmentation masks, extended by 11 pixel in all three directions to make sure the whole heart was captured in all scans of current and future cohorts. For training samples the bounding box was centered around the center of the manually segmented heart, while for testing samples the center of the predicted heart segmentation from the first step was used. The resulting sub-volumes, sized 384x384x80 pixels, were still too large to fit in the GPU memory. Hence, the scans were downsampled to 2.0x2.0x2.5mm resulting in an final input data size of 128x128x80 pixels. Although the output of this second step still had to be up-sampled, subsequent tests revealed high segmentation accuracy.

The model architecture employed in this step was the same as the one used in the first step. The initial learning rate was 10-5 and was dropped by 70% every 50 epochs with training achieving convergence after 1,000 epochs. To ensure the best segmentation results, we conducted an extensive hyperparameter selection process. After reaching a satisfying result, we froze the hyperparameters and trained the final model using a variation of cross validation to include all CT scans of the tuning dataset. Therefore we randomly split the full training cohort into a training and tuning set every 100 epochs, ensuring a different part of the full cohort being used for training. While this method made use of all training samples it still proved to effectively avoid overfitting on the training data. To further avoid overfitting, the data was augmented by applying rotations of ±35 degrees around the sagittal axis and translations of ±20px in the axial plane. The output of the model was again a probability mask which was thresholded by a factor of 0.99 to a binary mask and upsampled to the original scan size.

Data augmentation

Data augmentation has proven to be an efficient way to enhance the performance of a deep learning model but often comes at the cost of additional computational resources, and can introduce a significant increase in training time (Shorten and Khoshgoftaar, 2019). This tradeoff had to be considered carefully, as data augmentation of 3D data can be time intensive. Another drawback of data augmentation, especially critical in medical applications, is that it could yield to anatomy variations unlikely to be observed in real data.

For example, flipping CT images as a form of data augmentation could be detrimental, as the organs are not symmetrically placed within the human body. Moreover, all patients in our study underwent CT scanning in the same position. Therefore, we chose only moderate rotation and translation to simulate variations in patient positions on the CT scanner table, as well as variations in organ locations within the patient's bodies. To

avoid increasing the training time, scan augmentation was carried out using multi-core processing on the CPU in parallel to the model training on the GPU.

Statistical Analysis

The average heart size was calculated as the mean of the side lengths of the bounding boxes of all heart segmentations in the test cohorts. The heart alignment was calculated as the euclidean distance between the geometric centers of the bounding boxes of the automatic and manual heart segmentation. The statistical analysis and assessment of the segmentation accuracy was done using Python 2 with the MedPy package (Maier, 2015) for calculating the Dice coefficient (Dice, 1945), Jaccard coefficient (Jaccard, 1912), Hausdorff distance (Huttenlocher et al., 1993) and average symmetric surface distance.

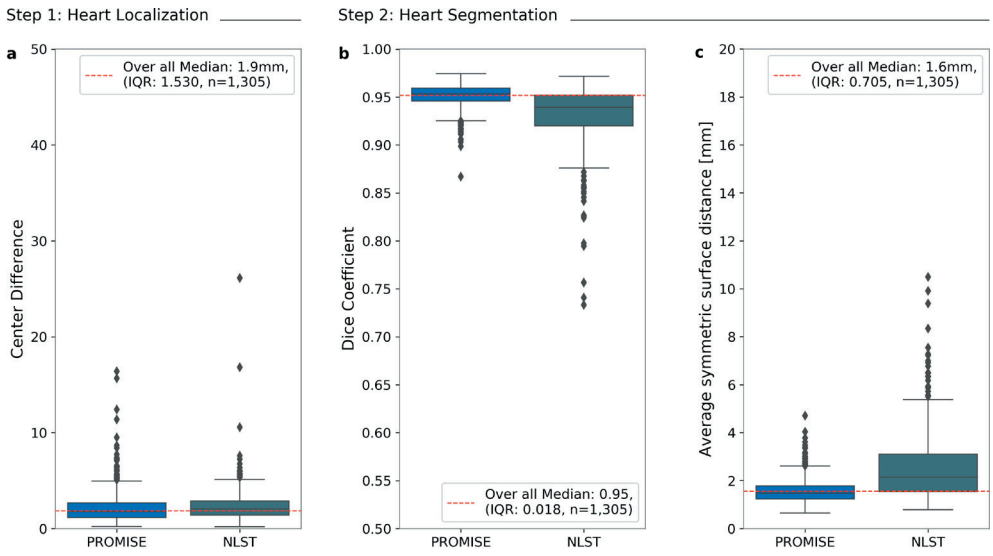


Figure 2. Performance results of the proposed deep learning system in cardiac ECG gated CT from the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) and in low dose chest screening CT from the National Lung Screening Trial (NLST). (a) Localization (Step 1) accuracy assessment showing the center difference between automatic and manual segmentations. Segmentation (Step 2) accuracy assessment using (b) the Dice coefficient and (c) the average symmetric surface distance.

Hard- and Software

The implementation and evaluation of the proposed deep learning system was carried out on a Linux workstation using Tensorflow-GPU (V1.14), Keras (V2.3.1), and NVIDIA CUDA (V10.1). We trained our models using two Intel Xeon CPUs and four NVIDIA RTX8000 GPUs. Training time for each model was approximately three days.

Results

A deep learning system for fully automatic heart segmentation was developed consisting of two consecutive steps to localize and subsequently segment the heart in a CT scan. To validate the performance of the deep learning system, we compared the automatic heart segmentations to manual segmentations from medical experts in two test cohorts in terms of localization and segmentation accuracy. The first test cohort included cardiac gated scans, while the second test cohort included non-gated low-dose chest screening scans (Figure 1).

Validation of the heart localization network (Step 1)

The first step of our deep learning system was to reliably locate the heart in a given CT scan independently of the scan size, resolution, and field of view. Our test cohort included 1,534 individuals from 226 participating medical sites. These scans differed especially in the number of slices (Scan height), resulting in big differences in the CT volume's sizes along the longitudinal axis, therefore making the heart localization step necessary. The proposed method localized the heart in the full test cohort with a median accuracy of 8±6mm compared to the gold standard of manual heart segmentations from medical experts. With an average heart bounding-box size in the test cohort of 119±17.5mm the calculated center lies within 6.7% of the heart size. Looking into the two test cohorts separately, the median center differences between automatic and manual heart segmentations were 6±4mm in cardiac gated scans from PROMISE and 13±9mm in low dose chest screening CTs from NLST (Figure 2a).

Validation of the heart segmentation network (Step 2)

After cropping the CT scans around the center of the heart determined in step 1, a second deep learning network was applied to compute a high resolution segmentation mask of the heart. Comparing automatic to manual segmentations, the average Dice coefficient in the complete test cohort was 0.94±0.03, the Jaccard coefficient was 0.88±0.04, and the average symmetric surface distance was 2.1±1.1mm. In more detail, in cardiac gated scans from PROMISE, the Dice coefficient was 0.94±0.02 and the Jaccard coefficient was 0.89±0.03 while the average symmetric surface distance was 1.9±0.7mm. In low dose chest screening CT from NLST the Dice coefficient was 0.92±0.03, the Jaccard coefficient was 0.86±0.07 and the average symmetric surface distance was 2.9±1.5mm (see Figure 2b and c). Next, we compared heart volume measurements of automatic and manual segmentations using the Spearman's correlation, which was very high for the complete test cohort (Mukaka, 2012) at 0.952 ($p<0.0001$). In the PROMISE test cohort the correlation was 0.964 ($p<0.0001$) and in the NLST test cohort it was 0.906 ($p<0.0001$) (see Figure 3). On our hardware, the proposed deep learning system needed 1.17 seconds per scan to

process, with 0.48 seconds per scan for the localization step and 0.69 seconds for the segmentation step.

For a visual assessment of the segmentation error in the two test cohorts, error-heatmaps were plotted and are shown in the Supplementary Figure S2. Three noticeable regions of segmentation uncertainty were located: The top of the heart, the bottom of the heart and the front of the heart, areas which are typically difficult to segment. Example segmentations can be seen in Figure 4 and Supplementary Figure S3. Furthermore, a complete summary of the results is shown in the Supplementary Table S1.

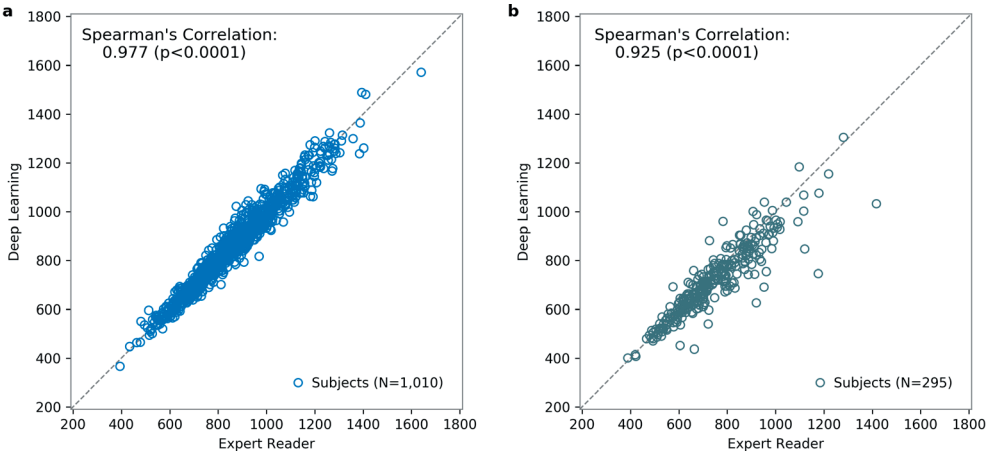


Figure 3. Volume comparison of automatic and manual heart segmentions (a) in cardiac ECG gated CT from the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) and (b) in low dose chest screening CT from the National Lung Screening Trial (NLST).

Discussion

In the presented investigation we developed a highly accurate and fast deep learning system to segment the heart in a variety of CT scans. The system was able to reliably locate the heart in dedicated cardiac CTs as well as full-body low-dose chest screening CTs and subsequently segmented the heart with high accuracy and precision. We assessed the performance of the system by calculating the heart center distance, the dice coefficient and average surface distance between manual and automatic heart segmentations.

A strength of this study was the size of the datasets for training, tuning and testing, which are one of the most important components for developing and evaluating deep learning systems. The large amount of high quality training cases enabled us to train networks which generalize well in large and independent test cohorts which were held out from training. To the best of our knowledge no other published study has used as many test cases acquired by this many medical sites as we did to prove the applicability of the proposed methods. Furthermore, the proposed network was able to accurately

segment the heart in ECG-gated cardiac CT and low-dose chest screening CT, with both methods not using contrast agents. Non-contrast enhanced scans are often routinely acquired in clinical practice to assess cardiovascular risk and are further reinforced by the current cholesterol guidelines to guide medical therapy in individuals with intermediate ASCVD score (Grundy et al., 2018). Furthermore they have less negative impact on the human body (No exposure to radiation dose of the contrast agent), save acquisition time, and are more cost effective, but on the other hand can make segmenting the heart a more challenging task. The high quality training data is also the reason for the great performance of our proposed deep learning system.

To show the generalizability of our deep learning system, we tested it on cardiac ECG-gated and non-gated scans. As expected, the segmentation accuracy in cardiac gated CT from PROMISE was higher than in the low-dose chest screening CTs from NLST. This can be explained due to the fact that the former typically have less motion artifacts and noise, resulting in more and distinct image features, and subsequently more precisely defined heart contours.

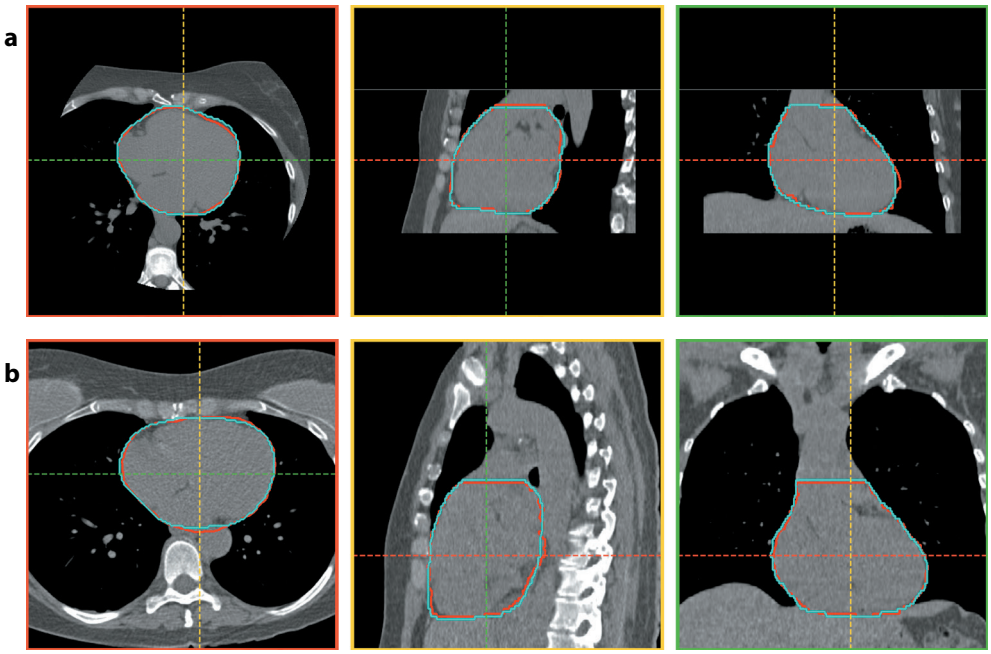


Figure 4. Comparison of the automatic and manual heart segmentation in an example case from the two test cohorts in axial (left), coronal (middle) and sagittal (right) direction. (a) Cardiac ECG-gated CT from the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE). The Dice coefficient in this case was 0.966, the average symmetric surface distance was 0.99mm and the distance between the segmentation centers was 2mm. (b) Low-dose chest screening CT from the National Lung Screening Trial (NLST). The Dice coefficient in this case was 0.964, the average symmetric surface distance was 1.1mm and the distance between the segmentation centers was 4mm. Note that the scan from the PROMISE cohort had a narrow field of view which was extended with air (Hounsfield unit: -1024) before processing.

Visual examination of the differences between automatic and manual heart segmentations revealed three common regions where errors occurred. The first, and apparently the most challenging area, was on top of the heart. This was expected as the upper end of the heart can not always be determined exactly and small variations in choosing the top slice consequently leads to high visible errors due to the slice spacing of 2.5mm of the CT scans. The second area of segmentation differences was at the bottom of the heart. This is most likely due to the fact that CT scans without contrast-agent lack image information in this area as the heart blends into surrounding organs. A precise segmentation in this area is often not possible even for most experienced cardiac radiologists. This problem also occurs when segmenting other organs in this area, for example the liver (Yang et al., 2017). The third region was located in the front of the heart where fat and muscle with low contrast and poor image information again make segmenting this part difficult. As confirmed by the mean surface distance values, the regions where the pipeline produces such errors are well-confined. The areas of segmentation differences are shown in Supplemental Figure S2. The difference in segmentations is higher in the NLST data than in PROMISE data. This is most likely due to the fact that NLST data is non-ECG-gated and hence contains more motion artefacts and noise, which make precise segmentation of the heart even more challenging.

We successfully used earlier development versions of the proposed system in several other studies. The first version of the system was used as a preprocessing step for a consecutive coronary artery calcium segmentation in over 20,000 CT scans (Zeleznik et al., 2021a). This early version was trained on 129 CTs only, resulting in lower performance than the later versions, but was already sufficiently accurate for the requirements of the project. We then extended the training cohort and used the deep learning system for fully automatic heart segmentation in 3,798 cardiac-gated CTs to assess the predictive value of whole heart volume for cardiac events (Foldyna et al., 2021). Furthermore we assessed the performance of the system in 5,677 planning CTs for the radiotherapy treatment of breast cancer patients (Zeleznik et al., 2021b). In this study, the network showed it's great potential for assisting dosimetrists at segmenting the heart, as the segmentation time significantly decreased, while intra-reader agreement increased, with accuracy being similar with and without deep learning assistance. While the earlier versions of the system already showed high segmentation accuracy, they all had high hardware requirements, most notably the use of four GPUs in parallel. The here presented model is now able to run on a single GPU as well as solely on a CPU, making it possible to be run on a wide range of computers. Furthermore we also improved the network structures and training process, leading to a better performance.

Comparing the performance of our network to previously published methods was not possible. Due to data privacy agreements, most studies did not provide their data publicly. Furthermore, differences in segmentation guidelines and task-specific quality requirements made testing on data from other studies less meaningful. Also worth

mentioning is that most existing publications focus on segmenting the whole heart plus sub-structures of the heart. This makes a comparison biased towards segmenting the whole heart only, as it is an easier and faster task. Therefore, we focused on comparing statistical outcomes rather than the computed segmentations. These circumstances as well as hardware differences also need to be considered when comparing the segmentation time per scan of the evaluated methods.

In 2017, the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) organized the Multi-Modality Whole Heart Segmentation Challenge (MM-WHS) (Zhuang et al., 2019), which provided 60 contrast enhanced cardiac CT acquired by two different medical sites, split into a training set (n=20) and testing set (n=40). Zhuang et al. evaluated the results of the challenge of the top ten participating teams. The accuracy of the whole heart segmentations of the evaluated teams was slightly below our results with a Dice between 80.6 and 90.8 and mean surface distance ranged from 4.8±13.6mm to 1.1±0.2mm.

Similar to our work, Habijan et al. (Habijan et al., 2019) published a study for automatic heart segmentation using two consecutive U-Nets. In addition, they also segmented sub-structures of the heart in 20 contrast enhanced cardiac CT from the 2017 MM-WHS, 15 for training and 5 for testing, reaching a Dice coefficient of 0.89 for the whole heart volume segmentation. The higher performance of our deep learning system shows the significance of a large training set including high quality data and independent validation data. This is also apparent in the work of Feng et al. (Feng et al., 2019), where they trained a deep convolutional neural network on 60 non-contrast and contrast enhanced CT scans from three different sites provided by the 2017 Auto-segmentation for Thoracic Radiation Treatment Planning AAPM challenge. While their model performed very well on the provided 12 test cases (Dice: 0.93), in an independent private test set including 30 cases the Dice decreased to 0.86, suggesting the network might have been overfitted on the training data. In general, the results of this challenge showed the great potential of deep learning based image segmentation. The winning team achieved a comparable accuracy for the heart segmentation with a Dice of 0.93 and a mean surface distance of 2.05mm, but these results are not directly comparable to our system performance as the test set of the challenge included 12 cases from which 3 were contrast enhanced scans (Yang et al., 2018).

In another promising work, Bui et Al. (2018) describe a fully automatic heart segmentation method based on a random walk model. They quantified their results in 58 patients and while they achieved great accuracy (Dice: 0.92), their approach was significantly slower with 1.88±0.71 minutes per scan compared to 1.17 seconds per scan of our network. This is a great example demonstrating how deep learning can improve image processing tasks, in terms of efficiency and speed, using code optimized to be processed on GPUs utilizing their massive parallelization potential.

Conclusions

The proposed deep learning system for fully automatic heart segmentation was developed utilizing high quality training, tuning and testing data, extensive medical expertise and state-of-the-art computational hardware. It showed great performance in two independent, large and diverse test cohorts promising great generalizability for new data and hence making it suitable for a vast array of medical and research applications. With this work we hope not only to show the potential of deep learning for fully automatic heart segmentation, but also to provide an accurate and reliable open source pre-trained model to the public without restriction and ultimately to achieve the transition of an experimental study to medical or scientific application.

CRediT author statement

Roman Zeleznik: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing;
Borek Foldyna: Conceptualization, Data curation, Supervision, Writing - original draft;
Jakob Weiss: Conceptualization, Data curation, Validation, Writing - original draft;
Parastou Eslami: Conceptualization, Data curation, Supervision;
Dennis Bontempi: Software, Validation, Writing - original draft;
Pamela S. Douglas: Data acquisition, review and editing;
Ramachandran S. Vasan: Data (CT scan) acquisition, Writing - review and editing;
Michael T. Lu: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review and editing;
Udo Hoffmann: Conceptualization, Funding acquisition, Project administration, Supervision;
Hugo J.W.L. Aerts: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing;

Acknowledgements

The authors thank the Framingham Heart Study, NCI, ACRIN, NLST, Prospective Multicenter Imaging Study for Evaluation of Chest Pain, and Rule Out Myocardial Infarction Using Computer Assisted Tomography II trial for access to trial data.
The authors acknowledge financial support from NIH (HA: NIH-USA U24CA194354, NIH-USA U01CA190234, NIH-USA U01CA209414, and NIH-USA R35CA22052; UH: NIH, 5R01-HL109711, NIH/NHLBI 5K24HL113128, NIH/NHLBI 5T32HL076136, NIH/NHLBI 5U01HL123339), the European Union - European Research Council (HA: 866504), as well as the German Research Foundation (DFG; TA: 1438/1-1 and WE: 6405/2-1), American Heart

Association Institute for Precision Cardiovascular Medicine (MTL: 18UNPG34030172), Fulbright Visiting Researcher Grant (E0583118), Rosztoczy Foundation Grant. The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the National Heart, Lung and Blood Institute.

Declaration of Competing Interest

The authors declare that they do not have any financial or non financial conflict of interests

Data and materials availability

NLST data is available upon request from the NCI (https://biometry.nci.nih.gov/cdas/nlst/) and the American College of Radiology Imaging Network (ACRIN, https://www.acrin.org/acrin-nlstbiorepository.aspx). The code of the deep learning system, as well as the trained model and statistical analysis will be shared to the public at publication of the manuscript. Furthermore, we will provide test data from a publicly available data set with manually and automatically computed segmentations and results. Although raw CT imaging data cannot be shared, measured results to replicate the statistical analysis will be made available. Published code and data will be available on the authors web page.

References

3D Slicer [WWW Document], n.d. URL http://www.slicer.org (accessed 7.13.20).

D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., Kannel, W.B., 2008. General Cardiovascular Risk Profile for Use in Primary Care. Circulation. https://doi.org/10.1161/circulationaha.107.699579

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology. https://doi.org/10.2307/1932409

Dimopoulos, K., Giannakoulas, G., Bendayan, I., Liodakis, E., Petraco, R., Diller, G.-P., Piepoli, M.F., Swan, L., Mullen, M., Best, N., Poole-Wilson, P.A., Francis, D.P., Rubens, M.B., Gatzoulis, M.A., 2013. Cardiothoracic ratio from postero-anterior chest radiographs: a simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. Int. J. Cardiol. 166, 453–457.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., 2017. Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. Communications in Computer and Information Science. https://doi.org/10.1007/978-3-319-60964-5_44

Douglas, P.S., Hoffmann, U., Lee, K.L., Mark, D.B., Al-Khalidi, H.R., Anstrom, K., Dolor, R.J., Kosinski, A., Krucoff, M.W., Mudrick, D.W., Patel, M.R., Picard, M.H., Udelson, J.E., Velazquez, E.J., Cooper, L., PROMISE investigators, 2014. PROspective Multicenter Imaging Study for Evaluation of chest pain: rationale and design of the PROMISE trial. Am. Heart J. 167, 796–803.e1.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. Nature 546, 686.

Feng, X., Qing, K., Tustison, N.J., Meyer, C.H., Chen, Q., 2019. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. Med. Phys. 46, 2169–2180.

Foldyna, B., Zeleznik, R., Eslami, P., Mayrhofer, T., Scholtz, J.-E., Ferencik, M., Bittner, D.O., Meyersohn, N.M., Puchner, S.B., Emami, H., Pellikka, P.A., Aerts, H.J.W.L., Douglas, P.S., Lu, M.T., Hoffmann, U., 2021. Small whole heart volume predicts cardiovascular events in patients with stable chest pain: insights from the PROMISE trial. Eur. Radiol. https://doi.org/10.1007/s00330-021-07695-2

Goeller, M., Achenbach, S., Marwan, M., Doris, M.K., Cadet, S., Commandeur, F., Chen, X., Slomka, P.J., Gransar, H., Cao, J.J., Wong, N.D., Albrecht, M.H., Rozanski, A., Tamarappoo, B.K., Berman, D.S., Dey, D., 2018. Epicardial adipose tissue density and volume are related to subclinical atherosclerosis, inflammation and major adverse cardiac events in asymptomatic subjects. J. Cardiovasc. Comput. Tomogr. 12, 67–73.

Grundy, S.M., Stone, N.J., Bailey, A.L., Beam, C., Birtcher, K.K., 2018. Guideline on the management of blood cholesterol: a report of the American College of Cardiology/American heart association Task force on clinical …. J. Am. Coll. Cardiol.

Habijan, M., Leventic, H., Galic, I., Babin, D., 2019. Whole Heart Segmentation from CT images Using 3D U-Net architecture. 2019 International Conference on Systems, Signals and Image Processing (IWSSIP). https://doi.org/10.1109/iwssip.2019.8787253

Hoffmann, U., Massaro, J.M., Fox, C.S., Manders, E., O'Donnell, C.J., 2008. Defining normal distributions of coronary artery calcium in women and men (from the Framingham Heart Study). Am. J. Cardiol. 102, 1136–41, 1141.e1.

Hosny, A., Parmar, C., Coroller, T.P., Grossmann, P., Zeleznik, R., Kumar, A., Bussink, J., Gillies, R.J., Mak, R.H., Aerts, H.J.W.L., 2018. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med. 15, e1002711.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.232073

Jaccard, P., 1912. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytologist. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Lu, M.T., Park, J., Ghemigian, K., Mayrhofer, T., Puchner, S.B., Liu, T., Fleg, J.L., Udelson, J.E., Truong, Q.A., Ferencik, M., Hoffmann, U., 2016. Epicardial and paracardial adipose tissue volume and attenuation - Association with high-risk coronary plaque on computed tomographic angiography in the ROMICAT II trial. Atherosclerosis 251, 47–54.

Mahbod, A., Chowdhury, M., Smedby, Ö., Wang, C., 2018. Automatic brain segmentation using artificial neural networks with shape context. Pattern Recognit. Lett. 101, 74-79.

Maier, O., 2015. MedPy.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.

Mohan, R., Barest, G., Brewster, L.J., Chui, C.S., Kutcher, G.J., Laughlin, J.S., Fuks, Z., 1988. A comprehensive three-dimensional radiation treatment planning system. International Journal of Radiation Oncology*Biology*Physics. https://doi.org/10.1016/s0360-3016(98)90033-5

Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., Bowling, M., 2017. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. Science 356, 508–513.

Mukaka, M.M., 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med. J. 24, 69–71.

National Lung Screening Trial Research Team, Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., Gatsonis, C., Marcus, P.M., Sicks, J.D., 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. 365, 395–409.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, pp. 234–241.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data 6, 60.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. Nature 550, 354–359.

Tong, G., Li, Y., Chen, H., Zhang, Q., Jiang, H., 2018. Improved U-NET network for pulmonary nodules segmentation. Optik. https://doi.org/10.1016/j.ijleo.2018.08.086

Um, H., Jiang, J., Thor, M., Rimner, A., Luo, L., Deasy, J.O., Veeraraghavan, H., 2020. Multiple resolution residual network for automatic thoracic organs-at-risk segmentation from CT. arXiv [eess.IV].

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M.L., Stolcke, A., Yu, D., Zweig, G., 2017. Toward Human Parity in Conversational Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. https://doi.org/10.1109/taslp.2017.2756440

Yang, D., Xu, D., Kevin Zhou, S., Georgescu, B., Chen, M., Grbic, S., Metaxas, D., Comaniciu, D., 2017. Automatic Liver Segmentation Using an Adversarial Image-to-Image Network. Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. https://doi.org/10.1007/978-3-319-66179-7_58

Yang, J., Veeraraghavan, H., Armato, S.G., 3rd, Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., Aljabar, P., Oliveira, B., van der Heyden, B., Zamdborg, L., Lam, D., Gooding, M., Sharp, G.C., 2018. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Med. Phys. 45, 4568–4581.

Zeleznik, R., Foldyna, B., Eslami, P., Weiss, J., Alexander, I., Taron, J., Parmar, C., Alvi, R.M., Banerji, D., Uno, M., Kikuchi, Y., Karady, J., Zhang, L., Scholtz, J.-E., Mayrhofer, T., Lyass, A., Mahoney, T.F., Massaro, J.M., Vasan, R.S., Douglas, P.S., Hoffmann, U., Lu, M.T., Aerts, H.J.W.L., 2021a. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nat. Commun. 12, 715.

Zeleznik, R., Weiss, J., Taron, J., Guthier, C., Bitterman, D.S., Hancox, C., Kann, B.H., Kim, D.W., Punglia, R.S., Bredfeldt, J., Foldyna, B., Eslami, P., Lu, M.T., Hoffmann, U., Mak, R., Aerts, H.J.W.L., 2021b. Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer. NPJ Digit Med 4, 43.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., Yang, X., Heng, P.-A., Mortazi, A., Bagci, U., Yang, G., Sun, C., Galisot, G., Ramel, J.-Y., Brouard, T., Tong, Q., Si, W., Liao, X., Zeng, G., Shi, Z., Zheng, G., Wang, C., MacGillivray, T., Newby, D., Rhode, K., Ourselin, S., Mohiaddin, R., Keegan, J., Firmin, D., Yang, G., 2019. Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge. Med. Image Anal. 58, 101537.

4

Chapter 4

Small whole heart volume predicts cardiovascular events in patients with stable chest pain: Insights from the PROMISE trial

Borek Foldyna, Roman Zeleznik, Parastou Eslami, Thomas Mayrhofer, Jan-Erik Scholtz, Maros Ferencik, Daniel O Bittner, Nandini M Meyersohn, Stefan B. Puchner, Hamed Emami, Patricia A. Pellikka, Hugo JWL Aerts, Pamela S Douglas, Michael T. Lu, Udo Hoffmann

Published in: European radiology (2021)

Abstract

Objectives

The size of the heart may predict major cardiovascular events (MACE) in patients with stable chest pain. We aimed to evaluate the prognostic value of 3D whole heart volume (WHV) derived from non-contrast cardiac computed tomography (CT).

Methods

Among participants randomized to the CT arm of the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE), we used deep learning to extract WHV, defined as volume of the pericardial sac. We compared the WHV across categories of cardiovascular risk factors and coronary artery disease (CAD) characteristics and determined the association of WHV with MACE (all-cause death, myocardial infarction, unstable angina; median follow-up: 26 months).

Results

In the 3,798 included patients (60.5±8.2 years; 51.5% women), the WHV was 351.9±57.6 cm^3/m^2. We found smaller WHV in no- or non-obstructive CAD, women, people with diabetes, sedentary lifestyle, and metabolic syndrome. Larger WHV was found in obstructive CAD, men, and increased atherosclerosis cardiovascular disease (ASCVD) risk score ($P<0.05$). In a time-to-event analysis, small WHV was associated with over 4.4-fold risk of MACE (HR (per one standard deviation) = 0.221; 95%CI: 0.068–0.721; $P=0.012$) independent of ASCVD risk score and CT-derived CAD characteristics. In patients with non-obstructive CAD, but not in those with no- or obstructive CAD, WHV increased the discriminatory capacity of ASCVD and CT-derived CAD characteristics significantly.

Conclusions

Small WHV may represent a novel imaging marker of MACE in stable chest pain. In particular, WHV may improve risk stratification in patients with non-obstructive CAD, a cohort with unmet need for better risk stratification.

Introduction

Cardiac computed tomography (CT) is increasingly used to exclude obstructive coronary artery disease in patients presenting with stable chest pain. According to the most recent European Society of Cardiology Guidelines, cardiac CT represents a first-line diagnostic method to assess cardiovascular (CV) risk in patients with chronic coronary syndromes, including those with stable chest pain [1]. While cardiac CT is a reliable method to exclude coronary artery disease (CAD) (negative predictive value ~99%) and to detect obstructive CAD, assessment of CV risk remains difficult, especially in those with non-obstructive disease [2]. Symptomatic patients with non-obstructive CAD, however, account for the majority of future CV events [3, 4], require advanced risk stratification, and are frequently referred to further testing.

CT-derived measures beyond stenosis assessment have revealed an additional prognostic value. For instance, elevated coronary artery calcium (CAC) or presence of high-risk plaque features (HRPF) on CT angiograms have been associated with increased risk of major adverse CV events (MACE) [3, 5, 6]. In addition to showing coronary arteries, cardiac CT has an advantage to image adjacent anatomical structures. Advanced CAD phenotyping, incorporating these structures, may leverage additional information and improve risk stratification. For example, epicardial adipose tissue, size of individual cardiac chambers, or CT-derived cardiac function have been related to adverse CV events beyond coronary stenosis and clinical risk factors [7–10].

Regarding heart morphology, the diameter of the heart on X-ray, and its proportion to thorax size (i.e., cardiothoracic ratio), are established measures of CV risk [11–14]. However, the prognostic value of CT-derived whole heart volume (WHV), a detailed 3D measure of heart size, available in all cardiac CT scans, has not been evaluated yet. Thus, this study's primary aim was to determine the association of WHV with MACE, adjusting for traditional measures of CV risk (i.e., atherosclerotic cardiovascular disease [ASCVD] risk score) and CAD characteristics on CT. In a final step, we performed a subgroup analysis across CAD categories (i.e., no-, non-obstructive, and obstructive CAD) and determined whether WHV had discriminatory capacity incremental to ASCVD risk score and CT-derived CAD characteristics.

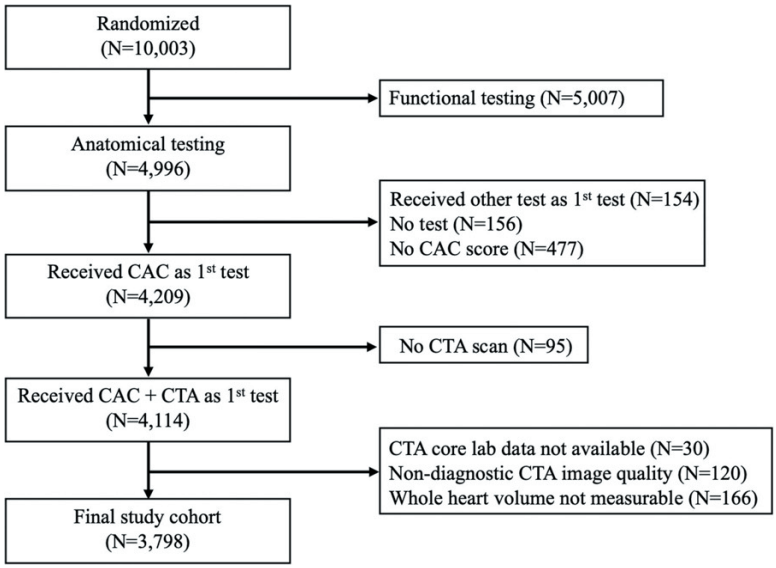


Figure 1. Consort diagram. *CAC:* coronary artery calcium; *CTA:* computed tomography angiography.

Methods

Study population and clinical characteristics

In this sub-study of the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) trial, we included patients who were randomized to anatomical testing and received non-contrast cardiac CT and contrast-enhanced coronary CT angiography (CTA). As per the PROMISE trial inclusion criteria, patients with known CAD or heart failure were not included. We excluded patients who received the first test other than CTA, did not undergo testing, received non-contrast CT only, or those with unavailable or non-diagnostic image data (**Consort diagram Figure 1**). Demographics and traditional CV risk factors were assessed with standard methods at the time of enrollment to the PROMISE trial[15]. Local and central institutional review boards approved the study, and all patients provided written informed consent.

Follow-up and the endpoints

All patients were followed for a median of two years. The primary endpoint was MACE, defined as a composite of all-cause mortality (CV + non CV death), non-fatal myocardial infarction (MI), and hospitalization for unstable angina (UA), as adjudicated by an independent committee [15].

WHV – definition and measurements

WHV (cm^3) was defined as volume of the pericardial sac, including all chambers (i.e., ventricles and atria), walls, and coronary arteries, but excluding epicardial fat (tissue ≤-45

HU). The cranial border was the axial slice at the level of the right mid pulmonary artery (**Figure 2**). To adjust for individual body size differences, we indexed the WHV by body surface area (BSA) (– cm^3/m^2) [16].

To decrease segmentation time, increase clinical feasibility, and standardize the measurement of WHV, we used a deep-learning system for the segmentation. The system consisted of two consecutive deep-learning networks of the U-Net architecture, to 1: localize and 2: segment the heart. The code was written in Python (v2.7) [17] using Tensorflow-GPU (v1.14) [18], Keras (v2.3.1) [19] with NVIDIA CUDA (v10.2) [20].

To ensure generalizability of the system, the training and tuning cohorts included 858 multi-center and multi-vendor CT scans from the Framingham Heart Study (FHS, n=628), the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE, n=130) and National Lung Screening Trial (NLST, n=100). Three experienced readers (BF, PE, JES) provided manually segmented hearts for the training (i.e., supervised learning). Here, the readers traced pericardial contours in axial images at 15 mm intervals and interpolated the space between the images using 3D Slicer (v.4.10) [21].

To further increase the segmentation accuracy, we automatically removed possibly incorrectly segmented lung tissue by excluding outer voxels with an attenuation < -400 HU. There was no manual correction of the automatic segmentations. The system accuracy was determined on an independent external validation dataset of 1,010 manually segmented hearts in PROMISE, revealing an excellent agreement (Dice coefficient: 0.94 ± 0.02).

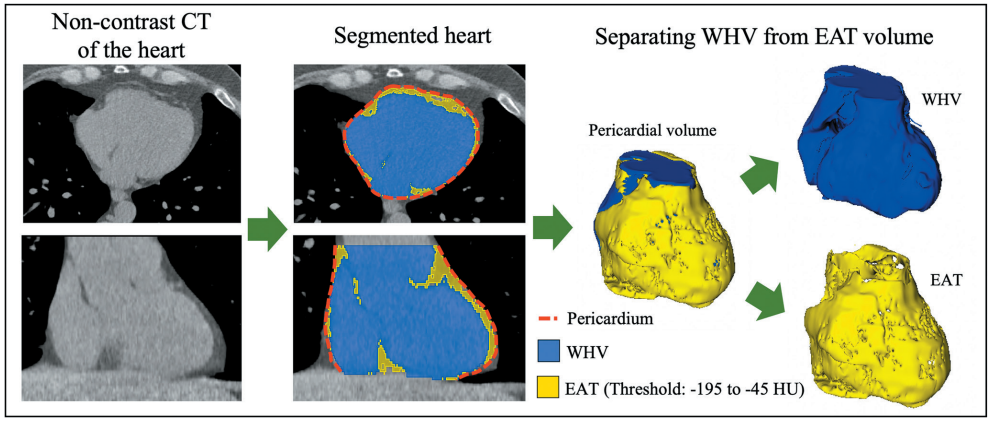


Figure 2. Measurements of WHV on non-contrast CT. Segmented hearts were derived from non-contrast cardiac CT images. The natural border was the pericardial sac (red-dotted line), and the segmentation ranged from the mid-right pulmonary artery (PA) to the most caudal part of the pericardial sac. To render WHV (blue), we subtracted the EAT volume (yellow), defined as fatty tissue with density thresholds of -195 to -45 HU. *CT:* computed tomography; *EAT:* epicardial adipose tissue; *HU:* Hounsfield units; *WHV:* whole heart volume.

CT-derived CAD characteristics

Experienced core lab readers measured CAC on non-contrast cardiac CT using the standard Agatston method [22]. Moreover, our core lab assessed all coronary arteries for the presence of CAD (non-obstructive: 1–69% and obstructive ≥70% maximal luminal narrowing in any coronary artery or ≥50% in the left main coronary artery) as well as the presence of HRPF as described elsewhere [5]. To determine CAD extent, accounting for plaque location and morphology, we calculated the Leaman score, an established tool to quantify total coronary atherosclerotic burden with information regarding localization, type of plaque, and degree of stenosis [23].

Statistical analysis

Continuous variables were expressed as mean ± standard deviation (SD) or median (interquartile range (IQR)) and categorical variables as frequencies and percentages. Differences of WHV across clinical characteristics were tested with Wilcoxon rank-sum test, while differences between categorical variables were tested with Fisher's exact test.

Univariate and multivariate Cox regressions were used to estimate the association of WHV with MACE. Results were reported as hazard ratios (HR) and 95% confidence intervals (CI). All regressions were stepwise adjusted for age, sex, ASCVD risk score, and Leaman score. Standard Kaplan-Meier survival curves incl. log-rank tests showed the differences in event-free survival across quintiles of WHV.

To test the incremental value of WHV, we evaluated whether the model fit increases significantly by adding WHV to the ASCVD risk and Leaman scores using the likelihood-ratio test for nested models. We also calculated receiver operator characteristic (ROC) curves to determine area under the curve (AUC) and estimate the increase of discriminatory capacity.

All analyses were performed in Stata 15.0 (College Station, TX), and two-sided *P*-values<0.05 were considered statistically significant.

Results

Study Population

Out of the 4,996 PROMISE patients randomized to anatomical testing, 3,798 fulfilled the inclusion criteria (**Figure 1**). The analytic cohort consisted of middle-aged (60.5±8.2 years), mostly overweight (median BMI: 30.3±5.8 kg/m^2), and predominantly male patients (58.5%) with intermediate CV risk (mean 10-year ASCVD risk score: 14.3±11.4%) (**Table 1)**. On coronary CTA, the mean Leaman score was 5.0±5.1, and HRPFs were present in 582/3,798 (15.3%) patients (. Over a median follow-up of 26.1 (18.0–34.4) months, 116/3,798 (3.1%) patients experienced MACE (MI: 21/116 (18.1%); death: 53/116 (45.7%); CV death: 30/116 (25.9%); UA: 46/116 (39.7%)).

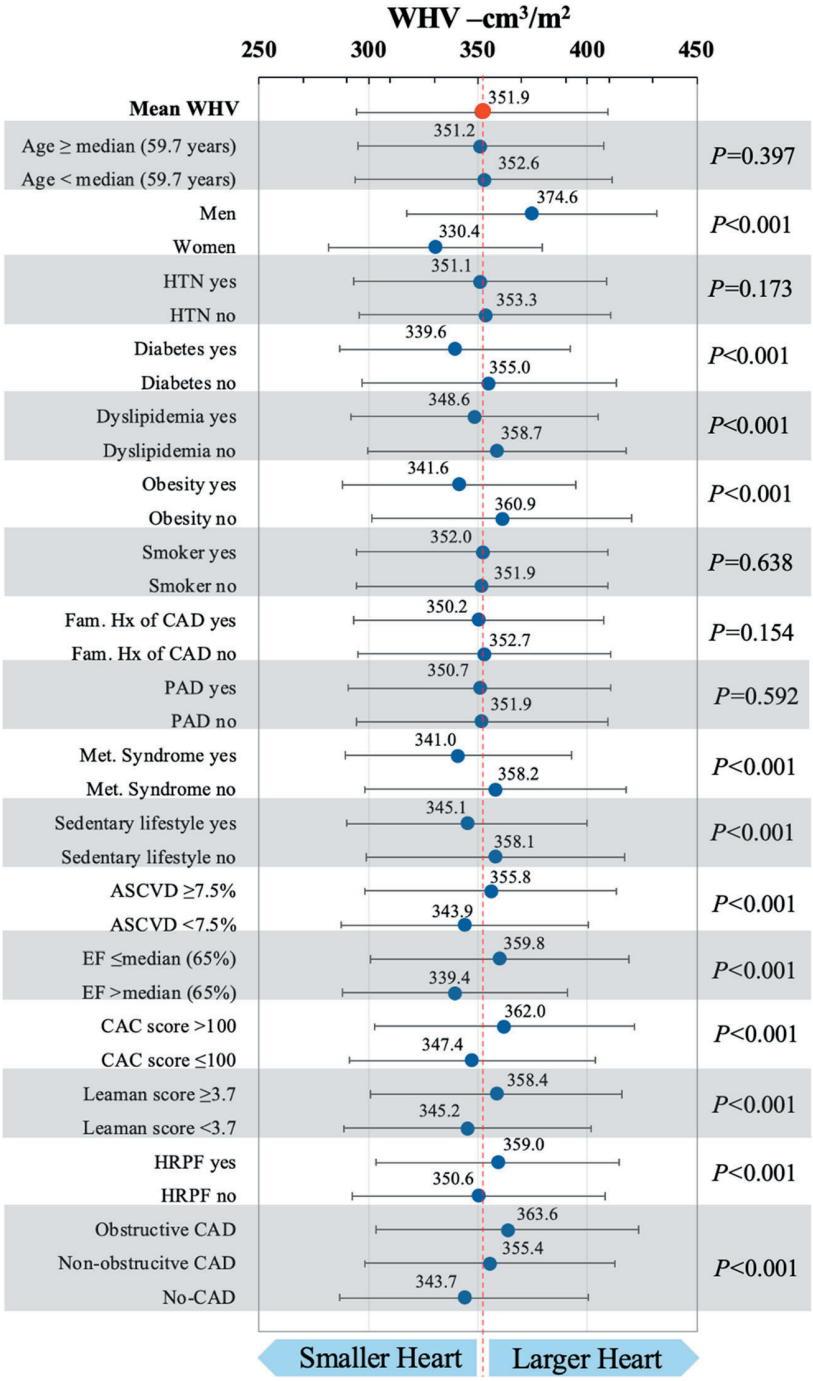


Figure 3. Whole heart volume, cardiovascular risk factors, and CAD on CT. Red-dotted line marks the **mean WHV** as reference (351.9cm3/m2). Bracketed lines represent standard deviations. *ASCVD:* atherosclerotic cardiovascular disease; *CAC:* coronary artery calcium; *CAD:* coronary artery disease; *EF:* ejection fraction; *HRPF:* high risk plaque features; *HTN:* arterial hypertension; *WHV:* whole heart volume; *PAD:* peripheral arterial disease.

Differences of WHV across categories of CV risk factors and CAD characteristics

The mean WHV was 351.9±57.6 cm^3/m^2 (range 100.6–746.1 cm^3/m^2). In general, men and those with increased ASCVD risk (≥7.5%) and advanced CAD (i.e., obstructive CAD, higher CAC score, HRPF present) presented with larger hearts (P<0.001 for all). On the other hand, women, patients with no- or non-obstructive CAD, obese patients, and those with metabolic syndrome, and sedentary lifestyle presented with significantly smaller hearts (P<0.05 for all). Separated by median age (59.7 years), WHV did not differ between younger and older patients (P=0.397). Individual WHV across categories of CV risk factors and CAD characteristics on CT are shown in **Figure 3**.

Table 1 Baseline characteristics

Mean ± SD or *n* (%)	All (*N* = 3798)	No MACE (*N* = 3682)	MACE (*N* = 116)	*p*
Demographics				
Age, years	60.5 ± 8.2	60.4 ± 8.2	63.0 ± 9.1	0.003
Women	1955 (51.5)	1905 (51.7)	50 (43.1)	0.073
Cardiovascular risk factors				
Hypertension	2441 (64.3)	2361 (64.1)	80 (69.0)	0.325
Diabetes mellitus	773 (20.4)	744 (20.1)	29 (25.0)	0.200
Dyslipidemia	2562 (67.5)	2486 (67.5)	76 (65.5)	0.687
BMI, kg/m^2	30.3 ± 5.8	30.3 ± 5.8	29.7 ± 5.7	0.255
Current or past smoker	1954 (51.5)	1878 (51.0)	76 (65.5)	0.002
Family history of premature (< 55 years) CAD	1258 (33.2)	1222 (33.3)	36 (31.0)	0.689
Any PAD	193 (5.1)	185 (5.0)	8 (6.9)	0.385
Metabolic syndrome	1379 (36.3)	1334 (36.2)	45 (38.8)	0.624
Sedentary lifestyle	1819 (48.0)	1747 (47.5)	72 (62.1)	0.002
Cardiovascular risk, %				
ASCVD risk score	14.3 ± 11.4	14.2 ± 11.3	20.2 ± 13.7	< 0.001
Relevant medication				
Beta-blocker	904 (24.8)	878 (24.8)	31 (27.4)	0.508
ACE inhibitor or ARB	1579 (43.4)	1529 (43.4)	50 (44.3)	0.848
Statin	1659 (45.6)	1611 (45.7)	48 (42.5)	0.565
Aspirin	1639 (45.0)	1589 (45.1)	50 (44.3)	0.924
Left ventricular EF*, %	64.6 ± 9.0	64.6 ± 8.9	65.2 ± 10.8	0.648
CAD on cardiac CT				
Coronary calcium score	20.2 (0.0–159.3)	18.2 (0.0–150.2)	146.4 (19.6–405.0)	< 0.001
Leaman score	3.7 (0.0–8.6)	3.7 (0.0–8.4)	8.0 (4.6–13.1)	< 0.001
No CAD	1297 (34.2)	1286 (34.9)	11 (9.5)	< 0.001
Non-obstructive CAD (1–69%)	2268 (59.7)	2192 (59.5)	76 (65.5)	
Obstructive CAD (≥ 70%)	233 (6.1)	204 (5.5)	29 (25.0)	

ACE, angiotensin-converting enzyme; *ARB*, angiotensin receptor blocker; *ASCVD*, atherosclerotic cardiovascular disease; *BMI*, body mass index; *CAD*, coronary artery disease; *CT*, computed tomography; *MACE*, major adverse cardiac events; *PAD*, peripheral arterial disease. *Available in a subgroup of 1815 patients. Values expressed as mean ± SD, median (IQR) or *N* (%)

Association of WHV with MACE

Patients who experienced MACE had smaller WHV compared to those without MACE (346.0±55.4 vs. 352.1±57.6 cm^3/m^2; p=0.005). In an age and sex-adjusted time-to-event analysis, we found that a decrease of WHV by one standard deviation was associated with over 4.4 times higher hazard of MACE (HR (per one standard deviation increase) = 0.225, 95%CI: 0.066–0.769, P=0.017). This association remained significant and at a similar magnitude after adjusting for the clinical ASCVD risk and Leaman score (adjusted

HR=0.221, 95%CI: 0.068–0.721, *P*=0.012). Additional results for raw WHV (i.e., not BSA-indexed) revealed similar results and are shown in **Table 2**. In a supplemental analysis, WHV remained significantly associated with MACE in a combined model adjusting for ASCVD, Leaman score, and CAC (BSA-indexed and ln-transformed WHV: HR=0.21, 95%CI: 0.066–0.710, *P*=0.011).

WHV across subgroups of CAD

On coronary CTA, 1,297 (34.2%), 2,268 (59.7%), and 233 (6.1%) patients presented with no-, non-obstructive, and obstructive CAD, respectively. Event rates differed significantly between patients without CAD and those with non-obstructive and obstructive disease (0.9% vs. 3.4% vs. 12.5%, respectively; log rank test, *P*<0.001). In patients with non-obstructive CAD, a decrease of WHV by one standard deviation was associated with 16.7 times higher hazard of MACE independent of ASCVD risk and Leaman score (adjusted HR=0.06, 95%CI: 0.013–0.269, *P*<0.001). However, there was no significant association between WHV and MACE in those with no- or obstructive CAD (*P*=0.146–0.853). **Table 3** provides results for raw WHV and BSA-adjusted WHV, which have shown similar results as the standardized WHV.

Table 2 Association of WHV with MACE in all patients with stable chest pain (*N* = 3798; MACE: *N* = 116)

	Adjusted for age and sex			Adjusted for ASCVD and Leaman score		
	HR	95% CI	*p*	HR	95% CI	*p*
WHV (absolute), mm^3	0.998	0.997–1.000	0.035	0.999	0.997–1.000	0.024
WHV (BSA-indexed), cm^3/m^2	0.996	0.992–0.999	0.022	0.996	0.992–0.999	0.018
WHV (BSA-indexed + *ln*-transformed)	0.225	0.066–0.769	0.017	0.221	0.068–0.721	0.012

Association of WHV with MACE (death, MI, or hospitalization for unstable angina). *ASCVD*, atherosclerotic cardiovascular disease; *BSA*, body surface area; *MACE*, major adverse cardiac events; *WHV*, whole heart volume

Non-obstructive CAD and WHV

In our cohort, the majority (n=76/116; 66%) of incident events occurred in the 2,268 patients with non-obstructive CAD. Among these, women presented with a slightly higher event rate compared to men (3.7% vs. 3.1%). Across quintiles of WHV, the MACE rate ranged between 2.3% and 5.4%, being nearly twice as high in the lowest quintile of WHV compared to Q2–5 (5.4% vs. 2.3–3.3%) In a time-to-event analysis, the lowest event-free survival was found in those with the lowest quintile of WHV (log-rank Q1 vs. Q2–5: *P*<0.001; **Figure 4)**. This association was further reflected in over two-fold higher hazard of MACE even after adjustment for ASCVD risk and Leaman score (adjusted HR (Q1 vs. Q2–5) =2.13; 95%CI: 1.29–3.51; *P*=0.003). In a sex-stratified analysis of patients with non-obstructive CAD, WHV showed an independent association with MACE in both women and men, being slightly stronger in men compared to women (Men: HR=0.064, 95%CI: 0.007–0.613, P=0.017 vs. Women: HR=0.080, 95%CI: 0.001–0.781, P=0.030 for BSA-indexed and ln-transformed WHV; **Supplemental Table S1**).

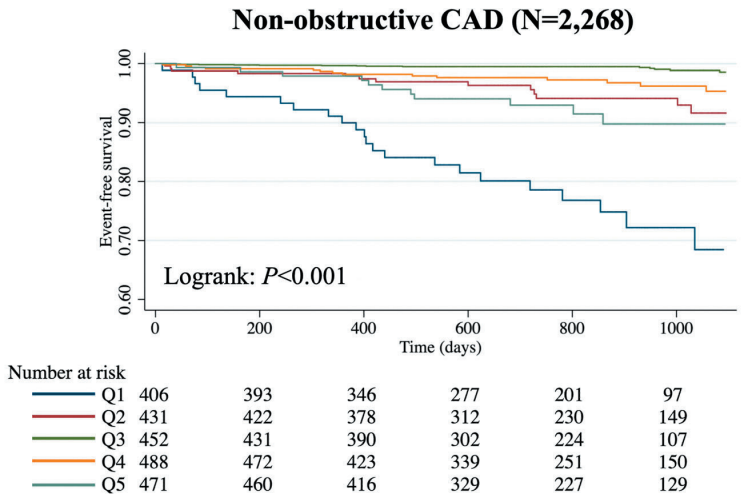


Figure 4. Quintiles of WHV and MACE in non-obstructive CAD. Significantly reduced event-free survival in patients with WHV in the first quintile (Q1) as compared to Q2–5 (log-rank results displayed as Q1 vs. Q2–5). All KM-curves were adjusted for ASCVD and Leaman score. *Q1–Q5:* quintiles of WHV. Kaplan-Meier curves for WHV in no- and obstructive CAD did not show significant results and are shown in **Supplemental Figure S1**.

Regarding event types, patients with small WHV experienced rather unspecific events, such as non-CV death or hospitalization for unstable angina, while those with larger hearts (Q2–5 of WHV) presented with more specific CV events, such as CV death or non-fatal myocardial infarction (**Figure 5).**

Table 3 Association of WHV with MACE stratified by CAD status on CTA

Adjustment	Age and sex			ASCVD and Leaman score		
	HR	95%CI	*P*	HR	95%CI	*P*
No CAD: 0% stenosis (*N* = 1297; MACE: *N* = 11)						
WHV (absolute), cm^3	1.002	0.998–1.007	0.327	1.002	0.999–1.006	0.255
WHV (BSA-indexed), cm^3/m^2	1.006	0.996–1.015	0.272	1.006	0.997–1.015	0.184
WHV (BSA-indexed + *ln*-transformed)	10.91	0.25–468.10	0.213	13.02	0.41–415.72	0.146
Non-obstructive CAD: 1–69% stenosis (*N* = 2268; MACE: *N* = 76)						
WHV (absolute), cm^3	0.997	0.995–0.999	0.003	0.997	0.996–0.999	0.001
WHV (BSA-indexed), cm^3/m^2	0.992	0.988–0.997	0.002	0.993	0.989–0.997	0.002
WHV (BSA-indexed + *ln*-transformed)	0.064	0.013–0.317	0.001	0.060	0.013–0.269	< 0.001
Obstructive CAD: ≥ 70% stenosis (*N* = 233; MACE: *N* = 29)						
WHV (absolute), cm^3	0.999	0.997–1.002	0.671	1.000	0.998–1.003	0.899
WHV (BSA-indexed), cm^3/m^2	0.998	0.991–1.005	0.623	0.999	0.993–1.006	0.803
WHV (BSA-indexed + *ln*-transformed)	0.55	0.04–7.01	0.648	0.80	0.08–8.14	0.853

Association of WHV with MACE (death, MI, or hospitalization for unstable angina) was driven by those with non-obstructive CAD. *ASCVD*, atherosclerotic cardiovascular disease; *BSA*, body surface area; *CAD*, coronary artery disease; *ln*, natural log; *MACE*, major adverse cardiac events; *WHV*, whole heart volume

Incremental value of WHV in non-obstructive CAD

In patients with stable chest pain and non-obstructive CAD, clinical parameters (i.e., ASCVD risk score), the CT-derived Leaman score, and WHV reached only a fair discriminatory capacity (AUC = 0.627, 0.599, and 0.589, respectively). While adding Leaman score to the ASCVD risk score did not lead to relevant changes of the AUC (0.627 vs. 0.627), addition of WHV to the model resulted in a statistically significant improvement of model fit (likelihood-ratio test (3 degrees of freedom): X^2=17.9; P<0.001). Correspondingly, the AUC increased by 4.6% reaching an AUC of 0.673 (**Figure 6**). We did not test for incremental value of WHV in patients with no- or obstructive CAD, since the initial tests (i.e., regressions) were negative.

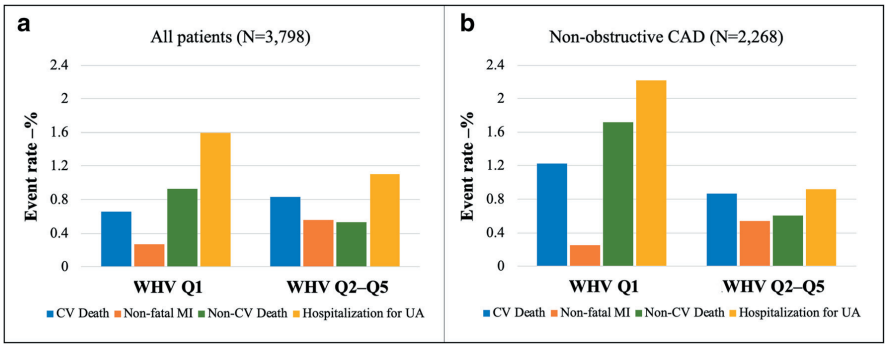
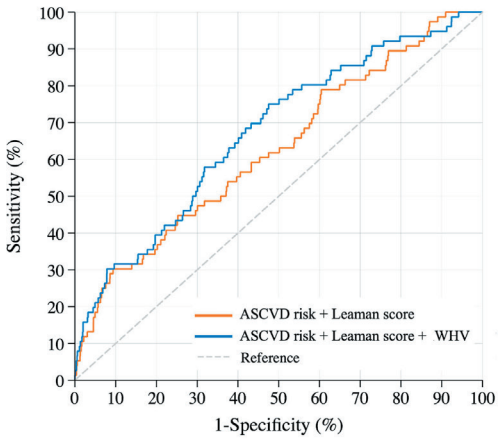


Figure 5. Event types and WHV. Patients with small WHV (Q1) presented with rather unspecific event types while those with higher WHV presented more frequently with specific cardiovascular events, an observation particularly seen in non-obstructive CAD. *CAD:* coronary artery disease; *CV:* cardiovascular; *MI:* myocardial infarction; *Q1-Q5:* quintiles of WHV; *UA:* unstable angina; *WHV:* whole heart volume.



Parameter	AUC	95% CI
ASCVD risk score	0.627	0.561–0.692
Leaman score	0.599	0.530–0.667
WHV	0.589	0.521–0.656
ASCVD risk + Leaman score	0.627	0.560–0.693
ASCVD risk + Leaman score + WHV	0.673	0.610–0.735

Figure 6. Incremental value of WHV in non-obstructive CAD. Addition of WHV to clinical risk factors and CT-derived CAD characteristics increased the discriminatory capacity significantly by 4.6%. Addition of WHV to the model resulted in a statistically significant improvement of model fit (X^2=17.9; P<0.001). *AUC:* area under the receiver operator characteristic curve; *ASCVD:* atherosclerotic cardiovascular disease; *WHV:* body surface area-indexed whole heart volume. To maintain readability, only curves for the composites are displayed here. Individual curves are available in **Supplemental Figure S2**.

Discussion

The primary finding of this study is that small WHV is an independent prognostic imaging marker of MACE among stable chest pain patients. This association is the strongest in those with non-obstructive CAD, a group of patients with the highest need for enhanced risk stratification. Moreover, in this group, WHV improves the discriminatory capacity of the traditional clinical CV risk factors and CTA-derived CAD characteristics.

Small hearts and MACE

Because our finding of an association of small WHV with MACE, especially in those with non-obstructive CAD, was independent of traditional CV risk factors and CAD burden, and that patients with small WHV experienced predominantly unspecific events, we suggest that the mechanism relating small WHV with MACE may not be directly linked to epicardial coronary atherosclerosis.

A clue, elucidating the potential pathophysiological/mechanistic link between small WHV and MACE, may be found in that small WHV was more frequent in women, people with diabetes, obese patients with metabolic syndrome, and sedentary lifestyle. This constellation, especially in the presence of non-obstructive CAD, has been described in the early stages of heart failure with preserved ejection fraction (HFpEF) [24–26]. Here, despite normal cardiac function, a combination of cardiometabolic disorder and non-obstructive CAD promotes myocardial fibrosis with concentric LV-remodeling [27, 28] and ultimately increased risk for MACE [27, 29]. PROMISE patients with small WHV had normal cardiac function based on prior definitions[30], including those with MACE.

Moreover, coronary microvascular dysfunction (CMD) and HFpEF are closely related, and a recent study found that 70–80% of patients with HFpEF also have CMD [31]. CMD and HFpEF also share clinical risk factors such as hypertension, diabetes, smoking, obesity, and chronic inflammatory disorders [32, 33], the majority of risk factors found in those with small WHV in our study. Our results add to the growing body of evidence, suggesting that CMD may be associated with non-obstructive CAD [32–35], and thus, may represent a potential link between non-obstructive CAD and HFpEF.

Given that PROMISE did not include patients with heart failure or known CAD (i.e., groups with often enlarged hearts), we hypothesize that the association between small WHV and MACE, may represent the left segment of a J-shaped relationship between WHV and MACE. This phenomenon has been described for other markers of CV risk, such as obesity [36]. However, this suggestion is hypothesis-generating and requires further investigation.

Large hearts and MACE

In our cohort, patients with more advanced CAD on cardiac CT (e.g., elevated CAC or Leaman score, obstructive CAD, or HRPF present) or, in general, elevated CV risk (i.e., ASCVD risk score ≥7.5%) presented with larger hearts. As expected, patients with larger hearts presented with rather typical CV events, such as CV death or non-fatal myocardial infarction. To some degree, these findings corroborate well-known associations of pathologically enlarged hearts, for example, those with clinical heart failure, dilatated or ischemic cardiomyopathy, and MACE [9, 14, 37–41]. It is crucial to understand that the PROMISE trial excluded patients with clinical signs of heart failure or known CAD, an aspect of selection that may explain why there was not a clear association between large hearts and MACE in our cohort.

Future perspectives

Future studies adding markers of structural and functional alterations of the heart (e.g., myocardial stiffness, interstitial collagen content, diastolic dysfunction, and strain), as well as CMD measures, are needed to test our hypothesis that small WHV relates to CMD and HFpEF in non-obstructive CAD. Moreover, studies of WHV in community-based populations, including those with enlarged hearts/heart failure, are needed to investigate the J-shaped relationship between WHV and MACE.

Clinical relevance

Our group and others have shown that non-obstructive CAD is related to an increased risk of MACE [3, 4]. In the PROMISE trial, non-obstructive CAD was associated with a three-fold increased risk for MACE compared to no CAD and accounted for the majority of events [3]. Thus, there is an unmet need for further risk stratification. Our study delivers a novel imaging marker that may improve risk stratification in this cohort at increased CV risk.

Limitations

Our study is a retrospective secondary analysis of a large randomized trial. Accordingly, our results are hypothesis-generating rather than confirmatory and need validation in large prospective cohorts. A comparison of WHV between patients with chest pain and normal WHV values was not possible since normal WHV has not been defined yet. Normal range of WHV will need to be derived from populations free of clinical symptoms. Despite the large scale of the PROMISE trial, the number of events is limited to provide reliable results in patient subgroups (e.g., quintiles of WHV in women and men with non-obstructive disease).

Conclusion

In stable chest pain patients, smaller WHV is an independent prognostic marker of MACE. Particularly in patients with non-obstructive CAD, small WHV may help to stratify CV risk beyond the traditional CV risk factors and CT-measures of CAD and may help to guide clinical management.

References

1. Knuuti J, Wijns W, Saraste A, et al 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromesThe Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). Eur Heart J. https://doi.org/10.1093/eurheartj/ehz425
2. Emami H, Takx RAP, Mayrhofer T, et al (2017) Non-obstructive CAD by coronary CTA improves risk stratification and allocation of statin therapy. JACC Cardiovasc Imaging 10:1031–1038. https://doi.org/10.1016/j.jcmg.2016.10.022
3. Hoffmann U, Ferencik M, Udelson JE, et al (2017) Prognostic Value of Noninvasive Cardiovascular Testing in Patients with Stable Chest Pain: Insights from the PROMISE Trial. Circulation. https://doi.org/10.1161/CIRCULATIONAHA.116.024360
4. Lin FY, Shaw LJ, Dunning AM, et al (2011) Mortality Risk in Symptomatic Patients With Nonobstructive Coronary Artery Disease: A Prospective 2-Center Study of 2,583 Patients Undergoing 64-Detector Row Coronary Computed Tomographic Angiography. J Am Coll Cardiol 58:510–519. https://doi.org/10.1016/j.jacc.2010.11.078
5. Ferencik M, Mayrhofer T, Bittner DO, et al (2018) Use of High-Risk Coronary Atherosclerotic Plaque Detection for Risk Stratification of Patients With Stable Chest Pain: A Secondary Analysis of the PROMISE Randomized Clinical Trial. JAMA Cardiol. https://doi.org/10.1001/jamacardio.2017.4973
6. Budoff Matthew J., Mayrhofer Thomas, Ferencik Maros, et al (2017) Prognostic Value of Coronary Artery Calcium in the PROMISE Study (Prospective Multicenter Imaging Study for Evaluation of Chest Pain). Circulation 136:1993–2005. https://doi.org/10.1161/CIRCULATIONAHA.117.030578
7. Lu MT, Park J, Ghemigian K, et al (2016) Epicardial and paracardial adipose tissue volume and attenuation – Association with high-risk coronary plaque on computed tomographic angiography in the ROMICAT II trial. Atherosclerosis 251:47–54. https://doi.org/10.1016/j.atherosclerosis.2016.05.033
8. Goeller M, Achenbach S, Marwan M, et al (2018) Epicardial adipose tissue density and volume are related to subclinical atherosclerosis, inflammation and major adverse cardiac events in asymptomatic subjects. J Cardiovasc Comput Tomogr 12:67–73. https://doi.org/10.1016/j.jcct.2017.11.007
9. Kizer JR, Bella JN, Palmieri V, et al (2006) Left atrial diameter as an independent predictor of first clinical cardiovascular events in middle-aged and elderly adults: the Strong Heart Study (SHS). Am Heart J 151:412–418
10. Bittencourt MS, Blankstein R, Mao S, et al (2016) Left ventricular area on non-contrast cardiac computed tomography as a predictor of incident heart failure – The Multi-Ethnic Study of Atherosclerosis. J Cardiovasc Comput Tomogr 10:500–506. https://doi.org/10.1016/j.jcct.2016.07.009

11. Dimopoulos K, Giannakoulas G, Bendayan I, et al (2013) Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. Int J Cardiol 166:453–457. https://doi.org/10.1016/j.ijcard.2011.10.125
12. Giamouzis G, Sui X, Love TE, et al (2008) A Propensity-Matched Study of the Association of Cardiothoracic Ratio With Morbidity and Mortality in Chronic Heart Failure††The Digitalis Investigation Group (DIG) study was conducted and supported by the NHLBI in collaboration with the DIG investigators. This report was prepared using a limited-access data set obtained by the NHLBI and does not necessarily reflect the opinions or views of the DIG study or the NHLBI. Am J Cardiol 101:343–347. https://doi.org/10.1016/j.amjcard.2007.08.039
13. Hemingway H, Shipley M, Christie D, Marmot M (1998) Cardiothoracic ratio and relative heart volume as predictors of coronary heart disease mortalityThe Whitehall study 25 year follow-up. Eur Heart J 19:859–869. https://doi.org/10.1053/euhj.1997.0862
14. Zaman MJS, Sanders J, Crook AM, et al (2007) Cardiothoracic ratio within the "normal" range independently predicts mortality in patients undergoing coronary angiography. Heart 93:491–494. https://doi.org/10.1136/hrt.2006.101238
15. Douglas PS, Hoffmann U, Lee KL, et al (2014) PROspective Multicenter Imaging Study for Evaluation of chest pain: rationale and design of the PROMISE trial. Am Heart J 167:796-803.e1. https://doi.org/10.1016/j.ahj.2014.03.003
16. Mosteller RD (1987) Simplified calculation of body-surface area. N Engl J Med 317:1098. https://doi.org/10.1056/NEJM198710223171717
17. Welcome to Python.org. In: Python.org. https://www.python.org/. Accessed 2 Dec 2020
18. TensorFlow. In: TensorFlow. https://www.tensorflow.org/. Accessed 2 Dec 2020
19. Keras: the Python deep learning API. https://keras.io/. Accessed 2 Dec 2020
20. (2017) CUDA Zone. In: NVIDIA Dev. https://developer.nvidia.com/cuda-zone. Accessed 2 Dec 2020
21. 3D Slicer. https://www.slicer.org/. Accessed 2 Dec 2020
22. Agatston AS, Janowitz WR, Hildner FJ, et al (1990) Quantification of coronary artery calcium using ultrafast computed tomography. J Am Coll Cardiol 15:827–832
23. de Araújo Gonçalves P, Garcia-Garcia HM, Dores H, et al (2013) Coronary computed tomography angiography-adapted Leaman score as a tool to noninvasively quantify total coronary atherosclerotic burden. Int J Cardiovasc Imaging 29:1575–1584. https://doi.org/10.1007/s10554-013-0232-8
24. Paulus WJ, Tschöpe C (2013) A Novel Paradigm for Heart Failure With Preserved Ejection Fraction: Comorbidities Drive Myocardial Dysfunction and Remodeling Through Coronary Microvascular Endothelial Inflammation. J Am Coll Cardiol 62:263-271. https://doi.org/10.1016/j.jacc.2013.02.092

25. Maaten JM ter, Damman K, Verhaar MC, et al (2016) Connecting heart failure with preserved ejection fraction and renal dysfunction: the role of endothelial dysfunction and inflammation. Eur J Heart Fail 18:588–598. https://doi.org/10.1002/ejhf.497
26. Kalogeropoulos A, Georgiopoulou V, Psaty BM, et al (2010) Inflammatory Markers and Incident Heart Failure Risk in Older Adults: The Health ABC (Health, Aging, and Body Composition) Study. J Am Coll Cardiol 55:2129–2137. https://doi.org/10.1016/j.jacc.2009.12.045
27. Velagaleti RS, Gona P, Pencina MJ, et al (2014) Left Ventricular Hypertrophy Patterns and Incidence of Heart Failure With Preserved Versus Reduced Ejection Fraction. Am J Cardiol 113:117–122. https://doi.org/10.1016/j.amjcard.2013.09.028
28. Shah RV, Abbasi SA, Heydari B, et al (2013) Insulin Resistance, Subclinical Left Ventricular Remodeling, and the Obesity Paradox: MESA (Multi-Ethnic Study of Atherosclerosis). J Am Coll Cardiol 61:1698–1706. https://doi.org/10.1016/j.jacc.2013.01.053
29. Pierdomenico SD, Lapenna D, Bucci A, et al (2004) Prognostic value of left ventricular concentric remodeling in uncomplicated mild hypertension. Am J Hypertens 17:1035–1039. https://doi.org/10.1016/j.amjhyper.2004.06.016
30. Fuchs A, Mejdahl MR, Kühl JT, et al (2016) Normal values of left ventricular mass and cardiac chamber volumes assessed by 320-detector computed tomography angiography in the Copenhagen General Population Study. Eur Heart J - Cardiovasc Imaging 17:1009-1017. https://doi.org/10.1093/ehjci/jev337
31. Shah SJ, Lam CSP, Svedlund S, et al (2018) Prevalence and correlates of coronary microvascular dysfunction in heart failure with preserved ejection fraction: PROMIS-HFpEF. Eur Heart J 39:3439–3450. https://doi.org/10.1093/eurheartj/ehy531
32. Camici PG, Crea F (2007) Coronary Microvascular Dysfunction. N Engl J Med 356:830–840. https://doi.org/10.1056/NEJMra061889
33. Tona F, Serra R, Di Ascenzo L, et al (2014) Systemic inflammation is related to coronary microvascular dysfunction in obese patients without obstructive coronary disease. Nutr Metab Cardiovasc Dis 24:447–453. https://doi.org/10.1016/j.numecd.2013.09.021
34. Crea F, Bairey Merz CN, Beltrame JF, et al (2017) The parallel tales of microvascular angina and heart failure with preserved ejection fraction: a paradigm shift. Eur Heart J 38:473–477. https://doi.org/10.1093/eurheartj/ehw461
35. Lee JF, Barrett-O'Keefe Z, Garten RS, et al (2016) Evidence of microvascular dysfunction in heart failure with preserved ejection fraction. Heart 102:278–284. https://doi.org/10.1136/heartjnl-2015-308403
36. Lavie CJ, McAuley PA, Church TS, et al (2014) Obesity and Cardiovascular Diseases: Implications Regarding Fitness, Fatness, and Severity in the Obesity Paradox. J Am Coll Cardiol 63:1345–1354. https://doi.org/10.1016/j.jacc.2014.01.022

37. Moller JE, Hillis GS, Oh JK, et al (2003) Left atrial volume: a powerful predictor of survival after acute myocardial infarction. Circulation 107:2207–2212
38. Vasan RS, Larson MG, Benjamin EJ, et al (1997) Left Ventricular Dilatation and the Risk of Congestive Heart Failure in People without Myocardial Infarction. N Engl J Med 336:1350–1355. https://doi.org/10.1056/NEJM199705083361903
39. Lauer MS, Evans JC, Levy D (1992) Prognostic implications of subclinical left ventricular dilatation and systolic dysfunction in men free of overt cardiovascular disease (the framingham heart study). Am J Cardiol 70:1180–1184. https://doi.org/10.1016/0002-9149(92)90052-Z
40. Raymond RJ, Hinderliter AL, Willis PW, et al (2002) Echocardiographic predictors of adverse outcomes in primary pulmonary hypertension. J Am Coll Cardiol 39:1214–1219. https://doi.org/10.1016/S0735-1097(02)01744-8
41. Sun JP, James KB, Sheng Yang X, et al (1997) Comparison of Mortality Rates and Progression of Left Ventricular Dysfunction in Patients With Idiopathic Dilated Cardiomyopathy and Dilated Versus Nondilated Right Ventricular Cavities. Am J Cardiol 80:1583–1587. https://doi.org/10.1016/S0002-9149(97)00780-7

5

Chapter 5

Epicardial adipose tissue in patients with stable chest pain: Insights from the PROMISE trial

Borek Foldyna, Roman Zeleznik, Parastou Eslami, Thomas Mayrhofer, Maros Ferencik, Daniel O Bittner, Nandini M Meyersohn, Stefan B. Puchner MD, Hamed Emami, Hugo JWL Aerts, Pamela S Douglas, Michael T Lu, Udo Hoffmann

Published in: Cardiovascular Imaging (2020)

Introduction

A growing body of evidence, primarily based on cohorts with low cardiovascular risk, suggests that epicardial adipose tissue (EAT), a metabolically active tissue surrounding coronary arteries, is associated with coronary artery disease (CAD) and adverse cardiac events [1,2]. However, little is known about the predictive value of EAT in symptomatic patients with elevated cardiovascular risk. We investigated the relationship between EAT, traditional cardiovascular risk factors, CAD characteristics, and incident adverse events in the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE), a symptomatic cohort with increased cardiovascular risk [3].

Methods

We included PROMISE patients who underwent cardiac computed tomography (CT) and followed them for a median two years for adverse cardiac events (composite of death, non-fatal myocardial infarction, and hospitalization for unstable angina) [3]. Local and central institutional review boards approved the study, and all patients provided written informed consent. Our dedicated core-lab measured EAT volume (indexed by body surface area –cm^3/m^2) and attenuation (HU) on non-contrast ECG-gated CT. We used a deep-learning-based system to segment EAT. The system consisted of two consecutive U-Nets to localize and segment pericardial sac, followed by an attenuation-based mask to render EAT (details provided in **Figure 1A**). Also, our core lab manually measured coronary artery calcium (CAC) score, CAD extent (segment involvement score; SIS), and presence of high-risk plaque features (HRPF: spotty calcium, positive remodeling, napkin ring sign, low attenuation plaque) using standard methods.

Our statistical analysis compared EAT volume between men and women, elderly (≥65 years) and younger patients, across categories of CAC, and in those with 10-year atherosclerotic cardiovascular disease (ASCVD) risk score ≥7.5% vs. <7.5%, SIS ≥4 vs. <4, and HRPF present vs. absent. In time to event analysis, we tested the association between EAT and events unadjusted and adjusting for ASCVD risk.

Results

In 3,948 patients (60.68.3 years; 51% women), mean EAT volume and attenuation were 57.5±22.0 cm^3/m^2 and -86.8±5.1 HU, respectively. Men presented with higher EAT volume and attenuation compared to women ($p<0.05$ for all, **Figure 1B**). The elderly and those with ASCVD ≥7.5% had more EAT but lower EAT attenuation compared to younger and those with ASCVD <7.5%.

Regarding CAD, EAT volume increased, and attenuation decreased proportionally to the CAC score ($p<0.001$ for both). Likewise, patients with SIS≥4 or HRPF had higher EAT

volume and lower attenuation compared to those with <4 stenotic segments or absent HRPF (p<0.001 for all, **Figure 1B**).

Overall, 128 (3.2%) patients experienced events during a median follow-up of 26.1 (18.0–34.4) months. Greater EAT volume was associated with a higher relative hazard of adverse events (HR per increase of 10 cm^3/m^2 EAT volume: 1.01; 95%CI: 1.00-1.15; p=0.036) in univariate analysis. However, this association attenuated after adjustment for the ASCVD risk (p=0.264). EAT attenuation was not associated with events (p=0.409).

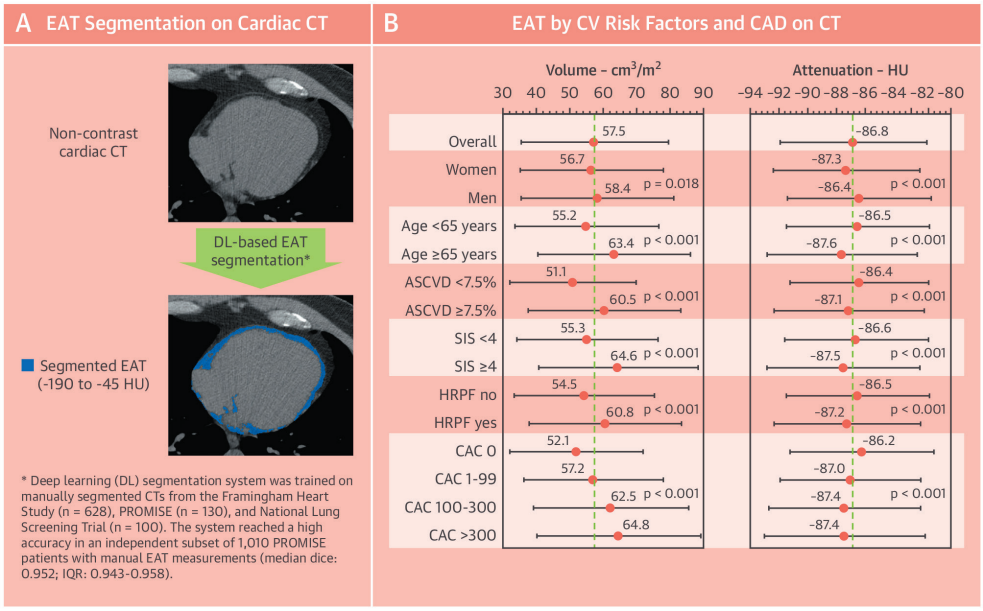


Figure 1. EAT on Cardiac CT Imaging. Epicardial adipose tissue (EAT) segmentation and EAT distribution across categories of cardiovascular risk factors and coronary artery disease (CAD) characteristics. ASCVD: Atherosclerosis Cardiovascular Disease risk score; CAC: coronary artery calcium; CT: computed tomography; EAT: epicardial adipose tissue; HRPF: high-risk plaque features; HU: Hounsfield unit; IQR: interquartile range; SIS: segment involvement score.

Discussion

In patients with stable chest pain and increased cardiovascular risk, we found higher EAT volume in males, the elderly, and those with increased cardiovascular risk and advanced CAD (i.e., higher CAC score, SIS, and HRPF). Nevertheless, these observations did not translate into an association between EAT volume and adverse events beyond traditional cardiovascular risk factors.

In accord with Mancio et al. (1), a meta-analysis with ~20,000 asymptomatic, low-risk subjects from mostly large longitudinal studies with long-term follow-up (e.g., Framingham Heart Study (FHS), Multi-Ethnic Study of Atherosclerosis (MESA), Heinz Nixdorf Recall (HNR) study, Early Identification of Subclinical Atherosclerosis by Non-invasive Imaging

Research (EISNER) study, and the Rotterdam study), our results demonstrated a strong relationship between EAT, cardiovascular risk factors, and extent of CAD. However, we did not find an independent association between EAT volume and adverse events.

In a symptomatic population, EAT may have low short-term prognostic utility due to many confounding elements such as clinical risk factors, presence of established CAD, and management strategy (e.g., medical therapy, revascularization) which are physician-dependent and may have varied between sites due to pragmatic design of the PROMISE trial. Moreover, PROMISE patients already had an indication for CTA, based on their symptoms and pretest probability. While longitudinal studies in asymptomatic cohorts have shown independent predictive prognostic value for deep-learning-derived EAT, independent of ASCVD risk score [2], the prognostic relationship may be attenuated for symptomatic patients undergoing CTA due to the reasons mentioned above.

We found no relationship between overall EAT attenuation and events. Thus, the cardiovascular risk may be better predicted by the local attenuation of EAT directly adjacent to the coronaries. As shown by others, pericoronary EAT attenuation has shown a strong association with cardiovascular events (4). The differences may be due to the local effects of EAT versus the global attenuation assessed in our study.

To conclude, our results show a limited short-term predictive value of EAT in symptomatic patients with increased cardiovascular risk. Further research is needed to identify factors influencing the relationship between EAT and adverse outcomes.

Financial Disclosure

Dr. Hoffmann received Research Grants from the National Institutes of Health (U01HL092040, U01HL092022), and Siemens Medical Solutions, Heart Flow Inc., and served as a consultant for Heart Flow. Dr. Lu reports consulting fees with PQBypass and a research grant from the Nvidia Corporation Academic Program. Dr. Lu is supported by grants from the American Heart Association Precision Medicine Institute 18UNPG34030172 and the Harvard University Center For AIDS Research NIH/NIAID 5P30AI060354-14. Dr. Ferencik reports receiving a grant from the American Heart Association. Dr. Picard received an unrelated stipend from the American Association of Echocardiography. The other authors have nothing to disclose.

References

1. Mancio J., Azevedo D., Saraiva F., et al. Epicardial adipose tissue volume assessed by computed tomography and coronary artery disease: a systematic review and meta-analysis. Eur Heart J - Cardiovasc Imaging 2018;19(5):490–7. Doi: 10.1093/ehjci/jex314.
2. Eisenberg Evann, McElhinney Priscilla A., Commandeur Frederic, et al. Deep Learning–Based Quantification of Epicardial Adipose Tissue Volume and Attenuation Predicts Major Adverse Cardiovascular Events in Asymptomatic Subjects. Circ Cardiovasc Imaging 2020;13(2):e009829. Doi: 10.1161/CIRCIMAGING.119.009829.
3. Douglas PS., Hoffmann U., Lee KL., et al. PROspective Multicenter Imaging Study for Evaluation of chest pain: rationale and design of the PROMISE trial. Am Heart J 2014;167(6):796-803.e1. Doi: 10.1016/j.ahj.2014.03.003.
4. Oikonomou EK., Marwan M., Desai MY., et al. Non-invasive detection of coronary inflammation using computed tomography and prediction of residual cardiovascular risk (the CRISP CT study): a post-hoc analysis of prospective outcome data. The Lancet 2018;392(10151):929–39. Doi: 10.1016/S0140-6736(18)31114-0.

6

Chapter 6

Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer

Roman Zeleznik*, Jakob Weiss*, Jana Taron, Christian Guthier, Danielle S. Bitterman, Cindy Hancox, Benjamin H. Kann, Daniel W. Kim, Rinaa S. Punglia, Jeremy Bredfeldt, Borek Foldyna, Parastou Eslami, Michael T. Lu, Udo Hoffmann, Raymond Mak, Hugo J.W.L. Aerts

Published in: npj Digital Medicine (2021)

Abstract

Although artificial intelligence algorithms are often developed and applied for narrow tasks, their implementation in other medical settings could help to improve patient care. Here we assess whether a deep learning system for volumetric heart segmentation on computed tomography (CT) scans developed in cardiovascular radiology can optimize treatment planning in radiation oncology. The system was trained using multi-center data (n=858) with manual heart segmentations provided by cardiovascular radiologists. Validation of the system was performed in an independent real-world dataset of 5,677 breast cancer patients treated with radiation therapy at the Dana-Farber/Brigham and Women's Cancer Center between 2008-2018. In a subset of 20 patients, the performance of the system was compared to eight radiation oncology experts by assessing segmentation time, agreement between experts, and accuracy with and without deep learning assistance. To compare the performance to segmentations used in the clinic, concordance and failures (defined as Dice<0.85) of the system were evaluated in the entire dataset. The system was successfully applied without retraining. With deep learning assistance, segmentation time significantly decreased (4.0 minutes [IQR 3.1-5.0] vs. 2.0 minutes [IQR 1.3-3.5]; p<0.001), and agreement increased (Dice 0.95 [IQR=0.02]; vs. 0.97 [IQR=0.02], p<0.001). Expert accuracy was similar with and without deep learning assistance (Dice 0.92 [IQR=0.02] vs. 0.92 [IQR=0.02]; p=0.48), and not significantly different from deep learning-only segmentations (Dice 0.92 [IQR=0.02]; p≥0.1). In comparison to real-world data, the system showed high concordance (Dice 0.89 [IQR=0.06]) across 5,677 patients and a significantly lower failure rate (p<0.001). These results suggest that deep learning algorithms can successfully be applied across medical specialties and improve clinical care beyond the original field of interest.

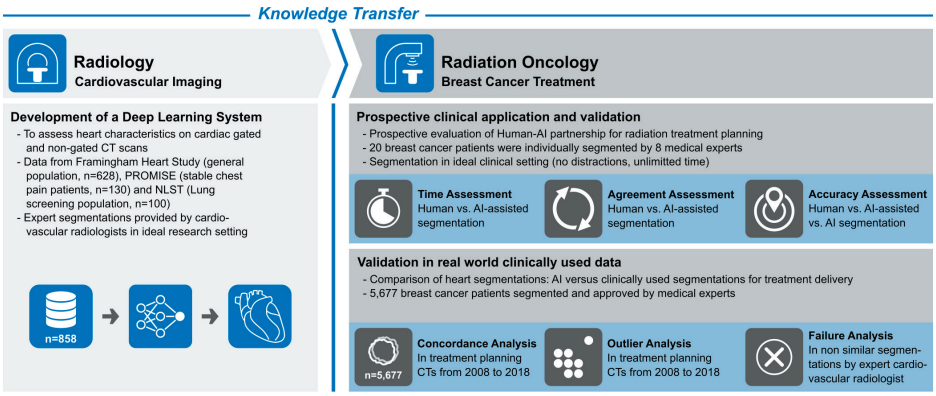


Figure 1. Study overview. A 3D deep learning system was developed in cardiovascular radiology using CT scans from distinct and well-established cohorts. For training, medical experts segmented the heart in cardiac gated and non-gated CT scans. This specialized knowledge embedded in the deep learning system was then transferred to radiation oncology and used to support treatment planning in patients with breast cancer.

Introduction

Medical knowledge is increasing exponentially with an estimated doubling every few months as of 2020[1]. While this has improved healthcare across the world[2], it is paralleled by increasingly specialized expert knowledge, which may be disproportionately distributed to high resource medical centers, thus increasing health care disparities[3]. Recent advances in artificial intelligence (AI), and deep learning in particular, offer a novel way to improve and automate complex tasks that up until now could only be performed by professionals[4]. Typically, deep learning applications are developed using labeled data generated by medical experts for domain-specific problems. As a result, this expert knowledge is encapsulated in the deep learning system, providing the opportunity to disseminate this highly skilled expertise across medical domains, institutions and countries, with the potential to optimize patient care and reducing knowledge and economic disparities in undersupplied settings.

One area that could benefit from this concept are imaging-related specialties, such as radiology and radiation oncology. While the former uses imaging studies primarily for diagnosis, the latter relies on the same information for organ and tumor targeting, treatment planning and delivery, and monitoring. An integral part of radiotherapy treatment planning is segmenting organs at risk in the radiation field on computed tomography (CT) scans[5]. If appropriate resources are available, this is done manually by trained experts who require considerable time and are prone to inter- and intra-observer variability. If time or knowledge are limited, this crucial step to ensure treatment quality and patient safety may be neglected. Therefore, automating and optimizing this process of organ at risk segmentation by deep learning could improve clinical care at high speed and low additional cost, especially in underprivileged healthcare settings[6].

Depending on the region of interest, different organs of varying complexity need to be segmented. Among those, the heart is of special interest as it is known that increasing radiation dose exposure to the organ is associated with future cardiac adverse events, such as coronary artery disease and heart failure[7,8]. Given their training, the highest anatomic expertise in cardiac imaging is likely found among cardiovascular radiologists, who focus on the diagnosis and monitoring cardiac-related diseases using dedicated image acquisition, reconstruction, and analysis techniques. Hence, disseminating this highly specific but narrow expert knowledge across medical domains and to institutions or countries with limited resources may enable more accurate treatment planning and measurement of cardiac radiation dose to optimize cardioprotective strategies in radiation oncology. This is of particular interest for patients with breast cancer as the heart and its substructures are in close proximity to the target area. Thus, reducing heart dose is of great importance to not harm the generally favourable outcomes of these patients.

Here, we investigate whether a deep learning system developed in cardiovascular radiology can be applied for radiation oncology treatment planning. The deep learning

system was developed for whole heart segmentation on input data provided by expert cardiovascular radiologists using dedicated, cardiac CT scans (see **Figure 1**). We then applied this system in an independent dataset with real-world segmentations of 5,677 patients with breast cancer to compare its performance to radiation oncology experts as well as to heart segmentations used in the clinic for treatment delivery. This study may serve as proof of principle to repurpose and leverage AI applications for optimizing patient care and reduce healthcare disparities across specialties, institutions, and countries.

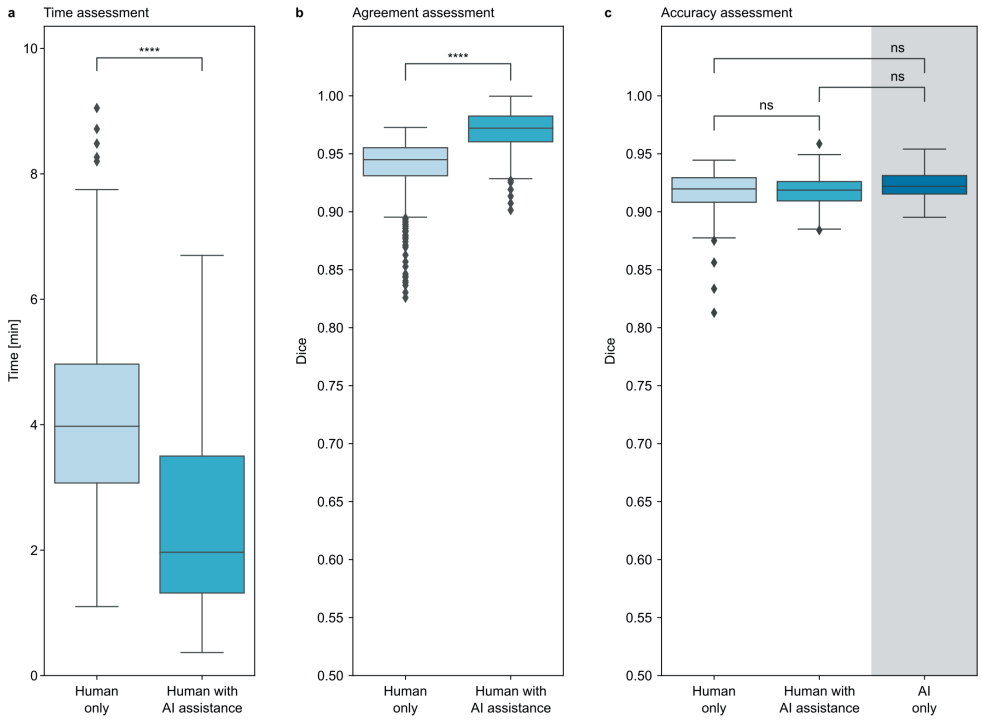


Figure 2. Comparison of human only, AI-assisted and AI only segmentation. In a prospective assessment, 8 radiation oncology experts individually segmented the heart in 20 breast cancer treatment CTs. In a subsequent session, the same patients were segmented again with AI assistance. **a**, The analysis shows that AI-assisted segmentation significantly reduces the time needed, (**b**) and agreement between medical experts significantly increases. **c**, Comparing the manual-only, AI-assisted and AI-only segmentations to the reference segmentations of a radiation oncology expert with several years of experience shows no significant differences in accuracy. Each box represents the interquartile range (IQR, 25th and 75th percentiles) and the centerline the median of the results. The whiskers represent minimum and maximum data points, excluding outliers. Outliers are defined as greater than the 75th percentile + 1.5ˣIQR and smaller than the 25th percentile - 1.5ˣIQR and are denoted as diamonds.

Results

Training, tuning, and testing the deep learning system

We trained and tuned the deep learning system with 757 ECG-gated cardiac CTs as well as 100 low dose chest screening CTs. The performance was tested in 1010 ECG-gated cardiac CTs and 296 low dose chest screening CTs. Manual segmentations were done under the supervision of cardiovascular radiologists at the Massachusetts General Hospital. The deep learning system achieved a median Dice of 0.95 (IQR=0.008) on the testing data.

Prospective validation of AI assistance in clinical setting

To evaluate the deep learning system for a clinical radiation oncology implementation, we compared the time needed to generate a clinically-acceptable segmentation without and with the assistance of the deep learning system, and found a significant reduction by 50% (median 4.0 minutes [IQR 3.1-5.0] vs. 2.0 minutes [IQR 1.3-3.5]; p<0.001) for the deep learning-assisted approach compared to the current manual clinical workflow (**Figure 2a**). At the same time, agreement of the segmentations significantly increased from a median Dice of 0.95 (IQR=0.02) for the manual segmentations to 0.97 (IQR=0.02) for the deep learning-assisted approach (p<0.001) (**Figure 2b**). Along with the changes in time and variation, accuracy analysis revealed no significant differences between the manual and deep learning-assisted segmentations (median Dice 0.92 [IQR=0.02] and 0.92 [IQR=0.02], respectively; p=0.50). Also, no significant differences were found between the deep learning-only segmentations (median Dice 0.92 [IQR=0.02]) and the manual as well as deep learning-assisted approach (p=0.2 and p=0.10, respectively) (**Figure 2c**). Additional results are provided in the **Supplementary Figure 2**.

Validation of performance in real-world, clinically-used data

In the subsequent assessment of the deep learning system in real-world data used in clinical practice, the automated whole heart segmentations showed a high concordance (median Dice of 0.89 [IQR=0.06]) with the clinically-used segmentations across the entire cohort of 5,677 breast cancer patients. In a per year analysis, the median Dice increased significantly from 0.85 (IQR=0.05) in 2008 to 0.91 (IQR=0.04) in 2018 (p<0.001). In parallel, the percentage of failure cases with a Dice below 0.85 decreased from 46.7% to 5.6%. An overview is given in **Figure 3a**.

The detailed failure analysis of cases with a Dice below 0.85 in the subset of patients treated between 2016 and 2018 comprised 299 patients. The ratings performed by the cardiovascular radiologist revealed a significantly higher segmentation accuracy for the deep learning system as compared to the manual, clinically-used segmentations (p<0.001). While the majority of deep learning segmentation were rated as excellent (62.2% vs. 15.7% for the clinical segmentations), most of the historical clinically-utilized segmentations were found to be of fair quality (70.6% vs. 35.5% for the deep learning system). Poor accuracy

was found for 13.7% of the clinical segmentations vs. 2.3% for the deep learning approach (**Figure 3b)**. Representative image examples are provided in **Figure 4**.

Discussion

In this study, we demonstrate that expert knowledge encapsulated in a deep learning system can be disseminated across medical domains to help optimize the treatment of patients with breast cancer in radiation oncology. The dissemination of domain-specific expert knowledge across disciplines in medicine has profound clinical implications. With the rapid and ongoing growth of knowledge across all medical specialties, no single discipline or individual can master the entire field of medicine beyond their expertise[1]. On the contrary, continued sub-specialisation and longer years of training lead to more narrow but highly skilled experts for particular fields or diseases. While this is beneficial if an expert is available onsite in resource-rich healthcare settings, the best possible care might not be deliverable to patients in low-resource areas[9]. In this context, expert knowledge encapsulated in deep learning systems developed and tested for specific tasks, but then re-purposed for different but related tasks in another specialty, institute or country, might be helpful to reduce knowledge and economic disparities, especially in undersupplied settings where such tasks might be neglected due to limited time or training. In such situations, an AI-mediated knowledge dissemination can create opportunities for human-AI partnerships to improve quality and safety in healthcare. Additionally, this approach maximizes the potential benefit of each expert annotated case, a particularly valuable aspect as deep learning tasks depend on such annotated data, and the current paucity of these data limits deep learning applications in medicine.

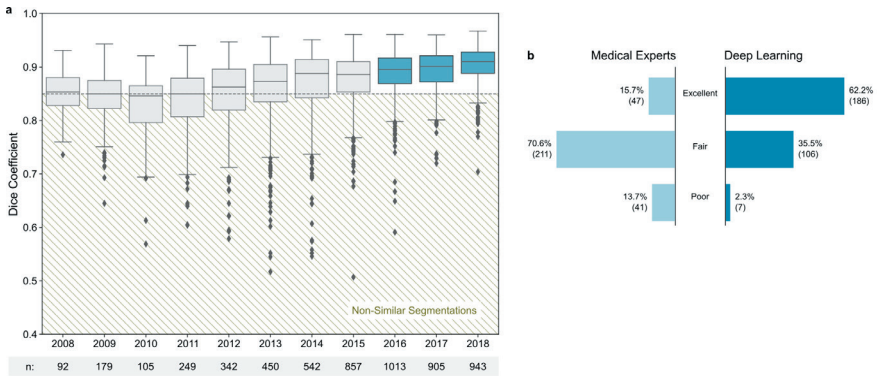


Figure 3. Similarity of manually and automatically generated segmentations. a, Dice coefficient between the AI framework and clinically approved heart segmentations in 5,677 scans acquired between 2008-2018. Non-similar segmentations with a dice coefficient below 0.85 (dashed line) were defined as failures. The boxes represent the interquartile range (IQR, 25th and 75th percentiles) and the centerlines the median of the results. The whiskers represent minimum and maximum data points, excluding outliers. Outliers are defined as greater than the 75th percentile + 1.5*IQR and smaller than the 25th percentile - 1.5*IQR and are denoted as diamonds. b, Results of qualitative segmentation accuracy assessment in cases defined as failures between 2016-2018 (n=299) by an expert cardiovascular radiologist. The results show significantly higher segmentation accuracy for AI as compared to radiation oncology experts (p-value < 0.0001).

In our prospective clinical assessment, we evaluated the potential of a human-AI partnership for heart segmentation as part of breast cancer radiation treatment planning. In a previously published study Tschandl et al.[10] showed how the human-AI relationship can improve image-based skin cancer diagnosis. In our study we found that the partnership between dosimetrist and AI facilitated the generation of highly accurate heart segmentations in a significantly shorter time and with a significantly higher concordance compared to the current clinical standard in a high resource medical center. At the same time, no differences in accuracy were observed. This is of considerable importance, as it helps to reduce labor-intensive manual work and could optimize quality while maintaining similar treatment standards[11]. Moreover, this is also an opportunity to improve the quality of care by reducing intra-reader and inter-reader variability both in radiology and radiation oncology[12,13], which persist despite standardized guidelines have been proposed to ensure quality control[14]. Most interestingly, when comparing the manual and deep learning-assisted segmentations to the deep learning-only segmentations, no differences in accuracy were found. This suggests that human input might not be necessary at all to generate segmentations of similar quality as the current clinical standard, thus suggesting the beginning of a paradigm shift in segmentation for radiotherapy treatment planning and the potential to implement this technique in undersupplied hospitals, in which organ at risk segmentation is not performed due to limited resources.

These results were emphasized in our assessment of the deep learning system in real-world, clinically-used data of 5,677 patients with breast cancer. Here, we could show a robust performance of the system without prior retraining. Although the median Dice was already high (0.85) in the subpopulation of patients treated in 2008, it significantly increased over the next decade to a median of 0.91. At the same time, the variance and number of patients with a Dice below 0.85 decreased. This is likely due to increased standardization of heart segmentations based on the 2016 RADCOMP guidelines, and recognition that heart dosimetry was intricately linked to radiotherapy toxicity and clinical outcomes[15]. In addition, it is of particular interest to gain a better understanding of failures before the potential implementation of a new deep learning system into clinical workflows. In our analysis of outlier cases with a Dice below 0.85, we found a significantly higher failure rate in the clinically-used segmentations as compared to the deep learning system (13.7% vs. 2.3%). This finding indicates that the current error rate in daily clinical practice could be significantly reduced by implementing the deep learning system for this heart segmentation task in radiotherapy planning. In addition, this may have implications for radiotherapy quality control by optimizing planning in order to minimize toxicity and enhance the therapeutic ratio. Moreover, creating a human-AI partnership for routine but clinically relevant tasks such as organ segmentation has the potential to fundamentally change and optimize clinical workflows[16]: 1) in high-resource centers by altering the role of medical experts from professionals spending substantial portions of their time manually generating segmentations to providing oversight of AI and quality control, while

freeing up more time for higher value responsibilities such as face-to-face interactions with patients and/or complex clinical decision-making and 2) in low-resource settings by introducing new treatment possibilities that are currently neglected but paramount for patient safety and quality of care.

The robust performance of the deep learning system is not only interesting from a clinical perspective, but also from a technical perspective[17]. In detail, we trained the deep learning system using images and segmentations from cardiovascular radiology and assessed the possibility to transfer this learned knowledge to radiation oncology and further studied the human-AI partnership. The main difference between images of the radiology training cohorts and images of the oncology testing cohorts was that the training cohorts included mostly cardiac ECG-gated CTs acquired during a breathhold interval to reduce cardiac and respiratory motion artifacts while the testing cohorts consisted solely of non-gated scans and many of them acquired during free-breathing. Segmentations in non-gated scans are typically less accurate due to motion artifacts. This applies to both manual and automatic segmentations and explains the small performance drop of our network from the training cohorts to our testing cohorts. In addition, acquisition and reconstruction protocols as well as scanners varied widely, however, that did not seem to have a major impact on performance. The difference between images from the training and testing set are shown in an example in the **Supplementary Figure 3**, indicating the different image acquisition and reconstruction techniques used in radiology and radiation oncology, respectively.

Although the input data in the current study was considerably different from the data used for development, no systematic failures were observed and differences in acquisition and reconstruction protocols did not affect the segmentation performance. This indicates the robustness of the deep learning system for potential applications in different clinical settings and beyond the primary intention of development. Additional data and patient baseline characteristics can be found in the **Supplementary Table 1**.

There are limitations to our study that need to be addressed. Time needed for segmenting the heart was self-recorded by the medical expert, which makes inaccurate measurements more likely than if they were taken by an independent person. Moreover, with the investigated AI system, only whole heart segmentations are possible although there is increasing evidence suggesting that cardiac substructures are of more importance and more closely linked to outcome and cardiac toxicity. Also, as the primary focus of this study was on deep learning-based expertise dissemination in a general setting, the analyses are lacking dose calculations and dedicated evaluations of treatment plans. In addition, as of now, only patients examined in supine position can be analyzed by the deep learning system given the input data used for training.

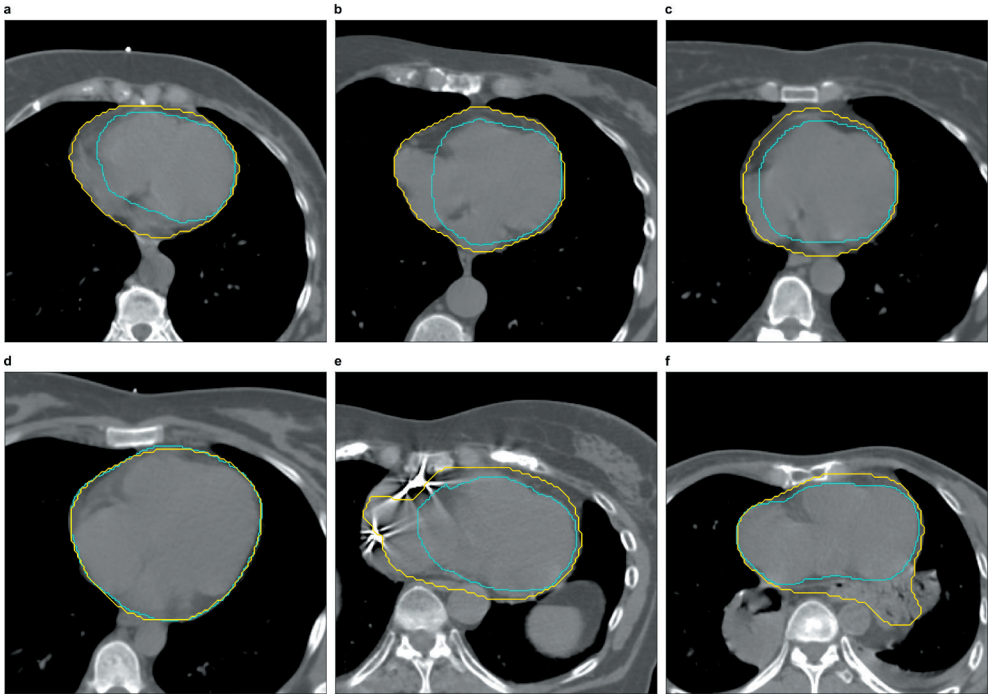


Figure 4. Segmentation accuracy of AI (Yellow) and manual (Cyan) segmentations. All manual segmentations were created by medical experts and approved by a radiation oncologist for treatment. Quality ratings (poor, fair, excellent) were made by a board-certified radiologist trained in cardiovascular imaging in a blinded pairwise fashion following the RTOG Breast Cancer Atlas. In **a, b, c,** AI was rated excellent whereas the clinically used segmentations revealed a poor accuracy (Dice: 0.811, 0.826 and 0.826 respectively). **d** depicts an example with excellent segmentation accuracy for AI and radiation oncology experts (Dice: 0.960). **e** shows poor accuracy for both, AI and radiation oncology experts (Dice: 0.773). In **f,** segmentation accuracy was rated poor for AI and fair for the radiation oncology expert segmentation (Dice: 0.833).

In conclusion, we demonstrated that expert knowledge encapsulated in a deep learning system can be disseminated across medical domains and institutes to optimize patient care beyond the intended narrow field of application. Furthermore, we demonstrated that the disseminated domain-specific expertise can be repurposed to substantially optimize the efficiency and quality of care in the investigated example of heart segmentation for breast cancer radiotherapy planning.

Methods

Study design and population

An overview of the study design is given in **Figure 1**. A search of the radiation oncology treatment planning system identified all breast cancer patients treated with radiotherapy in our institution's Department of Radiation Oncology between 2004-2018 (n=6,751). Exclusion criteria were: corrupted imaging data (n=380), missing/corrupted whole heart segmentations (n=499) and images of patients other than in supine position (n=195)

resulting in a final study cohort of 5,677 patients (**Supplementary Figure 1**). The study was conducted under a protocol approved by the Dana-Farber/Harvard Cancer Center institutional review board, which waved written informed consent. CT images for treatment planning were acquired following the institutional standards without administration of intravenous contrast agent and with and without breath holding. As the inclusion timeframe is over a decade, scanners as well as acquisition and reconstruction protocols varied widely, thus reducing the likelihood that the results are biased towards a single institution or a specific vendor, scanner or imaging technique, respectively. After reconstruction, images were transferred to the treatment planning system (Varian Eclipse, Varian Medical Systems, Palo Alto, California). All treatment plans and whole heart segmentation were created by trained medical experts following internal institution standards, and were in line with national guidelines as they became publicly-available starting in 2016 (e.g. RADCOMP Breast Cancer Atlas[15]). All heart segmentations were approved by an attending radiation oncologist for use in clinical treatment planning.

Development of AI system and domain transfer of expertise from cardiovascular radiology to radiation oncology

We developed a deep learning system, which is able to automatically localize and segment the heart from a given CT scan using expert knowledge from cardiovascular radiologists. Therefore the proposed system consists of two consecutive steps, each using a separate 3-dimensional deep learning model of the U-Net[18] architecture. In depth details of the system architecture, development, and application can be found in the Supplementary Methods (Supplementary Methods 1).

Prospective validation of AI assistance in radiation oncology

To prospectively investigate the potential of a human AI partnership, we assessed the performance of 8 trained medical experts (certified medical dosimetrists) responsible for radiation treatment planning by asking each expert to segment the whole heart using their typical clinical routine without and then with access to the deep learning system output. Measures of interest were 1) segmentation time, 2) agreement of the segmentations, defined as agreement between medical experts in the same patient, and 3) their anatomical accuracy, as outlined in RADCOMP Breast Cancer Atlas. For this assessment, 20 breast cancer patients were randomly selected from subjects treated in 2018. To avoid bias and ensure that the selected cases mirror a representative subset of the entire cohort, we calculated the dice coefficient between the AI segmentations and the clinically used segmentations before we started the trial with the dosimetrists. The mean dice was 0.90 (Std: 0.04) and the minimum and maximum dices were 0.78 and 0.94 respectively. As the network's performance was varying in the selected cases, we could assume that there was no bias in the selected subsample. Furthermore we used the parametric Welch's t-test and non-parametric Mann-Whitney U test to compare the Dice coefficients of the subset and

the full cohort. Both tests resulted in statistically not different dice coefficients (p=0.293 and p=0.153 respectively).

In a first segmentation session without distractions and no time limit, the medical experts were asked to segment the heart using the technique they would use in routine clinical care and recorded the time needed per patient. In a subsequent segmentation session 2 weeks later, the heart of the same 20 patients was pre-segmented with the deep learning system prior to the start of the session. The 8 medical experts were then asked to review and, where necessary, modify the deep learning segmentations until they were clinically acceptable for radiotherapy planning. Again, there were no other restrictions made and the time needed to modify the segmentations was self-recorded by each medical expert. The segmentations of a senior radiation oncologist with more than 16 years of experience in breast cancer treatment acquired in the same setting were used as reference standard for the medical experts as well as for the deep learning segmentations.

Validation in real-world data used for radiation treatment delivery

To investigate the application and robustness of the deep learning system in real-world clinically used data, we analyzed its performance across the entire study cohort using the historical, clinically-used segmentations as comparators. Based on a subjective review, a Dice ≥0.85 was arbitrarily defined as "similar segmentation". For quality control and to generate a better understanding for reasons of discordance between deep-learning and clinically-utilized heart segmentations, we manually analyzed cases considered as failures (Dice <0.85) in a subset of patients treated between 2016-2018 (n=299). This timeframe was chosen to explore failure rates in the most recently treated patients following the latest implemented guideline update[15]. A board-certified radiologist trained in cardiovascular imaging with 6 years of experience rated anatomical accuracy of the historical, manually created and clinically-used as well as the deep learning segmentations on a 3-point Likert scale (1=poor, 2=fair, 3=excellent). The reading session was performed in a pairwise fashion and blinded to the segmentation technique used.

Statistical analysis

All statistical analyses were performed in Python (V2.7). Data are presented as median and interquartile ranges (IQR). Similarity of manual and deep learning segmentations was measured using the Dice coefficient[19,20] with a smoothing factor of one. Furthermore, we calculated the symmetric surface distance and Hausdorff distance using the MedPy Python package (V0.4.4). For pairwise comparison a non-parametric Wilcoxon signed-rank test was performed due to violation of the normality assumption. To perform the parametric Welch's t-test and non-parametric Mann-Whitney U test we used the SciPy. stats Python package (V1.2.3). All p-values were two-sided and considered statistically significant below 0.05.

Data availability

The trained models are shared under the MIT license[21] at our webpage https://aim.hms.harvard.edu/DeepHeartRO. Due to privacy agreements with our institutions we can not share CT imaging or segmentation data. For that reason we provide test data from a publicly available data set with automatic heart segmentations.

Code availability

The full code of the deep learning system and statistical analysis is shared under the MIT license[21] at https://aim.hms.harvard.edu/DeepHeartRO.

Acknowledgements

The authors acknowledge financial support from the National Institutes of Health (HA: NIH-USA U24CA194354, NIH-USA U01CA190234, NIH-USA U01CA209414, and NIH-USA R35CA22052;
UH: NIH 5R01-HL109711, NIH/NHLBI 5K24HL113128, NIH/NHLBI 5T32HL076136, NIH/NHLBI 5U01HL123339) and the American Heart Association Institute for Precision Cardiovascular Medicine (MTL: 18UNPG34030172). JT is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – TA 1438/1-2. JW is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – WE 6405/2-1.

Competing Interests

RM discloses research grants from ViewRay, Inc. as well as consulting fees from ViewRay, Inc; AstraZeneca. All unrelated to this work. MTL reports consulting fees with PQBypass, research funding from MedImmune, and a GPU donation from the Nvidia Corporation Academic Program, all unrelated to this research. UH reports grants from HeartFlow, MedImmune, Siemens, Genentech, and the American College of Radiology Imaging Network and personal fees from the American Heart Association. HA reports consultancy fees and stock from Onc.AI, unrelated to this research. The remaining authors declare no competing interests.

Author contributions

Roman Zeleznik and Jakob Weiss equally contributed to this project. Detailed author contributions are as follows: Figures: Roman Zeleznik, Jakob Weiss; Code design, implementation and execution: Roman Zeleznik, Jakob Weiss; CT annotation: Jakob Weiss, Jana Taron, Christian Guthier, Danielle S. Bitterman, Daniel W. Kim, Benjamin H. Kann, Rinaa Sujata Punglia, Cindy Hancox; Study design: Roman Zeleznik, Jakob Weiss, Christian

Guthier, Danielle S. Bitterman, Michael Lu, Udo Hoffmann, Raymond Mak, Hugo J.W.L. Aerts; Training data preparation: Roman Zeleznik, Borek Foldyna, Parastou Eslami, Michael Lu, Udo Hoffmann, Raymond Mak, Hugo J.W.L. Aerts; Data analysis and interpretation: Roman Zeleznik, Jakob Weiss, Jana Taron, Christian Guthier, Raymond Mak, Hugo J.W.L. Aerts; Critical revision of the manuscript for important intellectual content: All authors; Statistical Analysis: Roman Zeleznik, Jakob Weiss; Study supervision: Raymond Mak, Hugo J.W.L. Aerts.

References

1. Densen, P. Challenges and opportunities facing medical education. *Trans. Am. Clin. Climatol. Assoc.* **122**, 48–58 (2011).
2. Craig, L. Service improvement in health care: a literature review. *British Journal of Nursing* vol. 27 893–896 (2018).
3. Hosny, A. & Hugo J W. Artificial intelligence for global health. *Science* vol. 366 955–956 (2019).
4. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat Biomed Eng* **2**, 719–731 (2018).
5. Mohan, R. *et al.* A comprehensive three-dimensional radiation treatment planning system. *International Journal of Radiation Oncology*Biology*Physics* vol. 15 481–495 (1988).
6. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
7. Gagliardi, G. *et al.* Radiation Dose–Volume Effects in the Heart. *International Journal of Radiation Oncology*Biology*Physics* vol. 76 S77–S85 (2010).
8. Darby, S. C. *et al.* Risk of ischemic heart disease in women after radiotherapy for breast cancer. *N. Engl. J. Med.* **368**, 987–998 (2013).
9. van Dis, J. MSJAMA. Where we live: health care in rural vs urban America. *JAMA* **287**, 108 (2002).
10. Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. *Nature Medicine* vol. 26 1229–1234 (2020).
11. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
12. Bauknecht, H.-C. *et al.* Intra- and Interobserver Variability of Linear and Volumetric Measurements of Brain Metastases Using Contrast-Enhanced Magnetic Resonance Imaging. *Investigative Radiology* vol. 45 49–56 (2010).
13. Steenbakkers, R. J. H. M. *et al.* Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis. *Int. J. Radiat. Oncol. Biol. Phys.* **64**, 435–448 (2006).
14. Huttin, C. The use of clinical guidelines to improve medical practice: main issues in the United States. *Int. J. Qual. Health Care* **9**, 207–214 (1997).
15. RADCOMP Breast Atlas. https://www.nrgoncology.org/About-Us/Center-for-Innovation-in-Radiation-Oncology/Breast/RADCOMP-Breast-Atlas.
16. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
17. Paschali, M., Conjeti, S., Navarro, F. & Navab, N. Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 493–501 (2018) doi:10.1007/978-3-030-00928-1_56.

18. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science* 234–241 (2015) doi:10.1007/978-3-319-24574-4_28.
19. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* vol. 26 297–302 (1945).
20. Sørensen, T. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. (1948).
21. The MIT License. https://opensource.org/licenses/MIT.

Chapter 7

Discussion and Conclusion

General Discussion

Deep learning has shown great success in several tasks, often being able to match and even surpass human performance. Especially for medical tasks, deep learning has the potential to support and improve research as well as clinical treatment, but, while there has been a large number of deep learning publications in the close past, real world deployment of such systems is still relatively sparse. Therefore, the goal of this thesis was not only to develop new deep learning based systems for medical applications by combining knowledge and experience of experts from various fields, but also to rigorously test their performance in large and diverse independent data sets to assess their performance for real world application.

One area where deep learning has been especially successful includes image related tasks, making it well suited for processing and evaluating CT scans from various medical fields. **Chapter 2** presented a deep learning based system for coronary calcium scoring in computed tomography to accurately stratify the risk for cardiovascular events across individuals from well-known NIH-sponsored observational cohorts and randomized controlled trials. The test data included over 20,000 individuals drawn from more than 200 medical sites, with prospective followup for cardiovascular events and death, which is the largest to date, to demonstrate the clinical value of automated calcium scoring. The test cohorts consisted of individuals from a primary prevention asymptomatic setting with non-gated chest CT (NLST)[56], as well as dedicated ECG-gated cardiac CT in stable (PROMISE)[57] and acute (ROMICAT-II)[58] chest pain setting. Varying health outcomes between datasets further reflected the diverse mix of asymptomatic and symptomatic individuals. Such diverse test data is essential. Before clinical introduction of automated systems can be considered, their generalizability has to be demonstrated by showing these systems are able to predict cardiovascular events of individuals across multiple clinical scenarios and perform robustly on data from multiple institutions.

Traditionally, coronary calcium scoring requires special software, manual measurement by trained experts and dedicated ECG-gated cardiac CT. As a consequence, the calcium score is often not reported on routine non-cardiac chest CT, despite the fact that calcium scores on non-gated CT have reasonably good agreement with dedicated calcium scoring CT[61,62]. The presented automatic calcium scoring system addresses this need by reliably and accurately extracting the calcium score from both cardiac CT and chest CT. Importantly, despite the known predictive value of coronary calcium for adverse cardiovascular events, dedicated coronary calcium scoring CT is not yet covered by Medicare and most US insurance companies, and for this reason, there is a great deal of interest in deriving the calcium score from routine chest CTs, which are far more common[32,63,64].

In summary, automated quantification of coronary calcium has the potential to improve clinical routine and population health. The deep learning calcium score

demonstrated robust risk prediction across all clinical scenarios with high correlation to human expert readers. Furthermore it showed robust test-retest reliability and calculated the calcium score in under two seconds. We shared our rigorously validated deep learning system with the public, allowing for accelerated adoption of these technologies by both academic and commercial entities.

Part of the presented deep learning system for calcium scoring was a robust and fast heart localization and segmentation in CTs. Several new research questions were the motivation to improve these two steps. **Chapter 3** presented a highly accurate and fast deep learning system to segment the heart in a variety of CT scans. The system was able to reliably locate the heart in dedicated cardiac CTs as well as full-body low-dose chest screening CTs and subsequently segmented the heart with high precision. Accuracy of the system was assessed by calculating the heart center distance as well as the dice coefficient and average surface distance between manual and automatic heart segmentations. A strength of this study was the size of the dataset for training and testing. The large amount of high quality training cases enabled the trained system to generalize well in large and independent test cohorts which were held out from training. To the best of our knowledge no other published study has used as many test cases acquired by this many medical sites as we did to prove the applicability of the proposed methods. Furthermore, the proposed network was able to accurately segment the heart in non-contrast enhanced cardiac ECG-gated CT and low-dose chest screening CT. Non-contrast enhanced scans are often routinely acquired in clinical practice to assess cardiovascular risk and are further reinforced by the current cholesterol guidelines to guide medical therapy in individuals with intermediate ASCVD score[31]. Furthermore, they have less adverse effects on the human body, save acquisition time, and are more cost effective, but on the other hand can make segmenting the heart a more challenging task.

As expected, the segmentation accuracy in cardiac ECG-gated CT from PROMISE was higher than in the low-dose chest screening CTs from NLST. This can be explained by the fact that the former CTs typically have less motion artifacts and noise, resulting in more and distinct image features, and subsequently more precisely defined segmentation contours.

Visual examination of the differences between automatic and manual heart segmentations revealed three common regions where errors occurred. The first, and apparently the most challenging area, was on top of the heart. This was expected as the upper end of the heart can not always be determined exactly and small variations in choosing the top slice consequently lead to high visible errors due to the slice spacing of 2.5mm of the CT scans. The second area of segmentation differences was at the bottom of the heart. This is most likely due to the fact that CT scans without contrast-agent lack image information in this area as the heart blends into the liver. A precise segmentation in this area is often not possible even for a senior cardiac radiologist and this problem also occurs when segmenting different organs in this area, such as the liver[65]. The third region

was located in the front of the heart where fat and muscle with low contrast and poor image information make segmenting this part difficult. As confirmed by the mean surface distance values, the regions where the pipeline produces such errors are well-confined. Compared to traditional implementations of automatic heart segmentation algorithms, the presented system shows that with a processing time of 1.17 seconds per scan, deep learning can profit and utilize the power of state of the art GPUs.

Utilizing the presented deep learning system for fully automatic heart segmentation, two studies were conducted to investigate the predictive value of various heart features for future cardiac events in almost 4,000 CT scans from the PROMISE trial. The study, presented in **Chapter 4**, showed that small whole heart volume is an independent prognostic imaging marker of major cardiovascular events among stable chest pain patients. Particularly in patients with non-obstructive coronary artery disease, small whole heart volume may help to stratify cardiovascular risk beyond the traditional cardiovascular risk factors and CT-measures of coronary artery disease and may help to guide clinical management. In the PROMISE trial, non-obstructive coronary artery disease was associated with a three-fold increased risk for major cardiovascular events compared to no coronary artery disease and accounted for the majority of events. Thus, there is an unmet need for further risk stratification. The presented novel imaging marker may improve risk stratification in this cohort of participants with increased cardiovascular risk. A second study used the deep learning system to segment epicardial adipose tissue in CT scans of patients with stable chest pain and increased cardiovascular risk, presented in **Chapter 5**. The study found higher epicardial adipose tissue volume in males, the elderly, and those with increased cardiovascular risk and advanced coronary artery disease. The results demonstrated a strong relationship between epicardial adipose tissue, cardiovascular risk factors, and extent of coronary artery disease. However, there was no independent association between epicardial adipose tissue volume and adverse events. Thus, the cardiovascular risk may be better predicted by the local attenuation of epicardial adipose tissue directly adjacent to the coronaries.

Another area where the heart is segmented in the daily clinical routine is radiotherapy treatment planning for breast cancer patients. The study presented in **Chapter 6** demonstrated that a deep learning system for automatic heart segmentation, developed in cardiovascular radiology can optimize treatment planning in radiation oncology. The study showed that expert knowledge encapsulated in a deep learning system can be disseminated across medical domains to help optimize the treatment of patients with breast cancer in radiation oncology. Additionally, this approach maximizes the potential benefit of each expert annotated case, a particularly valuable aspect as deep learning tasks depend on such annotated data, and the current paucity of these data limits deep learning applications in medicine.

Remarkably, although the data in this study was considerably different from the data used for the development of the deep learning system, no systematic failures were

observed and differences in acquisition and reconstruction protocols did not affect the segmentation performance. The main difference between images of the radiology training cohorts and images of the oncology testing cohorts was that the training cohorts included mostly cardiac ECG-gated CTs acquired during a breathhold interval to reduce cardiac and respiratory motion artifacts while the testing cohorts consisted solely of non-gated scans and many of them acquired during free-breathing. In addition, acquisition and reconstruction protocols as well as scanners varied widely, however, that did not seem to have a major impact on performance. The results indicated the robustness of the deep learning system for potential applications in different clinical settings and beyond the primary intention of development.

The study found that the partnership between dosimetrist and AI facilitated the generation of highly accurate heart segmentations in a significantly shorter time and with a significantly higher concordance compared to the current clinical standard in a high resource medical center. At the same time, no differences in accuracy were observed. This is of considerable importance, as it helps to reduce labor-intensive manual work and could optimize quality while maintaining similar treatment standards[15]. These results were emphasized in a further assessment of the deep learning system in real-world, clinically-used data of 5,677 patients with breast cancer. Here, the deep learning system showed a robust performance without prior retraining. In addition, it is of particular interest to gain a better understanding of failures before the potential implementation of a new deep learning system into clinical workflows. In an analysis of outlier cases with a low Dice was conducted, finding a significantly higher failure rate in the clinically-used segmentations as compared to the deep learning system. This finding indicates that the current error rate in daily clinical practice could be significantly reduced by implementing the deep learning system for this heart segmentation task in radiotherapy planning.

Future perspective

Deep learning is still a relatively new technology and its application in medical fields to date is relatively sparse. Before new methods can be implemented in clinical routines several hurdles have to be overcome. One important step is to show the robustness and generalizability of these methods by evaluating them in large and diverse test sets, which represents a major focus of the presented studies in this thesis. However, clinical cohorts are often biased towards specific ethnicities or other anatomical or social characteristics. For example, the majority of the individuals included in the datasets used in this thesis were predominantly non-Hispanic whites[66,56,67,68]. Therefore, it is of crucial importance to further assess the performance of the presented deep learning systems on racial and ethnic diverse populations in future studies[69].

There are also technical challenges that need to be solved in the future. Currently available deep learning frameworks are complex and their application is challenging

as it requires basic knowledge about software development. Easier to use methods are needed to bring deep learning to a broader audience, to further increase the application and acceptance of these new methods. The presented studies in this thesis contributed to this development by making all code and trained models publicly available. Immediate next steps have to focus on reducing hardware requirements for the currently presented deep learning methods (e.g., reduce the requirement of 4 GPUs to 1 GPU or even no GPU at all) and to containerize the code (e.g., using docker containers with GPU support), making it easier to run and apply.

Another major area that needs improvement in the future is data availability. As training deep learning models heavily relies on large amounts of good quality data, the access to such data gives research groups an edge over competition. Hence, data has become the capital of research groups, and data sharing is rarely seen with publications. On the other hand, sharing clinical data is often not possible, due to privacy requirements. As deep learning methods are applied in real world clinical routines, would need to reliably work in institutions around the world, on data from different machines and created with different methods, parameters and settings, large and diverse training cohorts are of essence. Here, future collaborations of groups around the world are needed to unite their knowledge, data and efforts to improve existing deep learning methods and find new solutions for medical problems.

There are several next steps for the presented studies of this thesis to further improve their performance and application. On the one hand, the methods will be used in new studies to process large datasets for future research. On the other hand, the presented methods will be used in clinical trials, investigating their real world performance and clinical impact, and ultimately driving further improvements while accelerating medical research and clinical treatment.

Conclusion

In this thesis, several novel deep learning approaches for medical applications were presented and rigorously tested in large, distinctive, and independent datasets to show their robustness and generalizability. The presented studies illustrated the predictive value of deep learning based risk predictions, highlighted the benefit of deep learning in clinical studies, and demonstrated that expert knowledge encapsulated in a deep learning system can be disseminated across medical domains and institutes.

References

1. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
2. Moravčík, M. *et al.* DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* **356**, 508–513 (2017).
3. Xiong, W. *et al.* Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 25 2410–2423 (2017).
4. Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
5. Machinery, C. Computing machinery and intelligence-AM Turing. *Mind* **59**, 433 (1950).
6. Russell, S. & Norvig, P. Artificial intelligence: a modern approach. (2002).
7. Langley, P. The changing science of machine learning. *Mach. Learn.* **82**, 275–279 (2011).
8. Ivakhnenko, A. G. & Lapa, V. G. *Cybernetics and Forecasting Techniques*. (Elsevier Science, 1968).
9. Ivakhnenko, A. G. Polynomial Theory of Complex Systems. *IEEE Trans. Syst. Man Cybern.* **SMC-1**, 364–378 (1971).
10. Dechter, R. Learning while searching in constraint-satisfaction-problems. in *ProceedingsAAAI'86* 178–183 (American Asso-ciation for Artificial Intelligence, 1986).
11. Parkhi, O. M., Vedaldi, A. & Zisserman, A. Deep face recognition. (2015).
12. Amodei, D. *et al.* Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. in *Proceedings of The 33rd International Conference on Machine Learning* (eds. Balcan, M. F. & Weinberger, K. Q.) vol. 48 173–182 (PMLR, 2016).
13. Otter, D. W., Medina, J. R. & Kalita, J. K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans Neural Netw Learn Syst* **32**, 604–624 (2021).
14. Rao, Q. & Frtunikj, J. Deep learning for self-driving cars. *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems* (2018) doi:10.1145/3194085.3194087.
15. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
16. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
17. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
18. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

19. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).
20. Dong, H., Yang, G., Liu, F., Mo, Y. & Guo, Y. Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. *Communications in Computer and Information Science* 506–517 (2017) doi:10.1007/978-3-319-60964-5_44.
21. Tong, G., Li, Y., Chen, H., Zhang, Q. & Jiang, H. Improved U-NET network for pulmonary nodules segmentation. *Optik* vol. 174 460–469 (2018).
22. Mohan, R. *et al.* A comprehensive three-dimensional radiation treatment planning system. *International Journal of Radiation Oncology*Biology*Physics* vol. 15 481–495 (1988).
23. Dimopoulos, K. *et al.* Cardiothoracic ratio from postero-anterior chest radiographs: a simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. *Int. J. Cardiol.* **166**, 453–457 (2013).
24. Um, H. *et al.* Multiple resolution residual network for automatic thoracic organs-at-risk segmentation from CT. *arXiv [eess.IV]* (2020).
25. Mahbod, A., Chowdhury, M., Smedby, Ö. & Wang, C. Automatic brain segmentation using artificial neural networks with shape context. *Pattern Recognit. Lett.* **101**, 74–79 (2018).
26. Zhuang, X. *et al.* Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge. *Med. Image Anal.* **58**, 101537 (2019).
27. Wilkins, E. *et al. European Cardiovascular Disease Statistics 2017*. (2017).
28. Writing Group Members *et al.* Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. *Circulation* **133**, e38–360 (2016).
29. Agatston, A. S. *et al.* Quantification of coronary artery calcium using ultrafast computed tomography. *J. Am. Coll. Cardiol.* **15**, 827–832 (1990).
30. Thanassoulis, G. *et al.* A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. *Circ. Cardiovasc. Genet.* **5**, 113–121 (2012).
31. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. *Journal of the American College of Cardiology* vol. 73 3168–3209 (2019).
32. Hecht, H. S. *et al.* 2016 SCCT/STR guidelines for coronary artery calcium scoring of noncontrast noncardiac chest CT scans: A report of the Society of Cardiovascular Computed Tomography and Society of Thoracic Radiology. *J. Thorac. Imaging* **32**, W54–W66 (2017).
33. Patel, M. R. *et al.* Correction to: ACC/AATS/AHA/ASE/ASNC/SCAI/SCCT/STS 2017 Appropriate Use Criteria for Coronary Revascularization in Patients With Stable Ischemic Heart Disease. *J. Nucl. Cardiol.* **25**, 2191–2192 (2018).

34. Raff, G. L. *et al.* SCCT guidelines on the use of coronary computed tomographic angiography for patients presenting with acute chest pain to the emergency department: A Report of the Society of Cardiovascular Computed Tomography Guidelines Committee. *Journal of Cardiovascular Computed Tomography* vol. 8 254–271 (2014).
35. Ravenel, J. G. & Nance, J. W. Coronary artery calcification in lung cancer screening. *Translational Lung Cancer Research* vol. 7 361–367 (2018).
36. Gupta, A. *et al.* The Identification of Calcified Coronary Plaque Is Associated With Initiation and Continuation of Pharmacological and Lifestyle Preventive Therapies: A Systematic Review and Meta-Analysis. *JACC Cardiovasc. Imaging* **10**, 833–842 (2017).
37. Budoff, M. J. *et al.* Prognostic Value of Coronary Artery Calcium in the PROMISE Study (Prospective Multicenter Imaging Study for Evaluation of Chest Pain). *Circulation* **136**, 1993–2005 (2017).
38. Goeller, M. *et al.* Epicardial adipose tissue density and volume are related to subclinical atherosclerosis, inflammation and major adverse cardiac events in asymptomatic subjects. *J. Cardiovasc. Comput. Tomogr.* **12**, 67–73 (2018).
39. Knuuti, J. *et al.* 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. *Eur. Heart J.* **41**, 407–477 (2020).
40. Emami, H. *et al.* Nonobstructive Coronary Artery Disease by Coronary CT Angiography Improves Risk Stratification and Allocation of Statin Therapy. *JACC: Cardiovascular Imaging* vol. 10 1031–1038 (2017).
41. Hoffmann, U. *et al.* Prognostic Value of Noninvasive Cardiovascular Testing in Patients With Stable Chest Pain: Insights From the PROMISE Trial (Prospective Multicenter Imaging Study for Evaluation of Chest Pain). *Circulation* **135**, 2320–2332 (2017).
42. Lin, F. Y. *et al.* Mortality risk in symptomatic patients with nonobstructive coronary artery disease: a prospective 2-center study of 2,583 patients undergoing 64-detector row coronary computed tomographic angiography. *J. Am. Coll. Cardiol.* **58**, 510–519 (2011).
43. Giamouzis, G. *et al.* A propensity-matched study of the association of cardiothoracic ratio with morbidity and mortality in chronic heart failure. *Am. J. Cardiol.* **101**, 343–347 (2008).
44. Hemingway, H., Shipley, M., Christie, D. & Marmot, M. Cardiothoracic ratio and relative heart volume as predictors of coronary heart disease mortality. The Whitehall study 25 year follow-up. *Eur. Heart J.* **19**, 859–869 (1998).
45. Pierdomenico, S. D., Pierdomenico, A. M., Neri, M. & Cuccurullo, F. Epicardial adipose tissue and metabolic syndrome in hypertensive patients with normal body weight and waist circumference. *Am. J. Hypertens.* **24**, 1245–1249 (2011).
46. Hatem, S. N. & Sanders, P. Epicardial adipose tissue and atrial fibrillation. *Cardiovasc. Res.* **102**, 205–213 (2014).

47. Mancio, J. *et al.* Epicardial adipose tissue volume assessed by computed tomography and coronary artery disease: a systematic review and meta-analysis. *Eur. Heart J. Cardiovasc. Imaging* **19**, 490–497 (2018).
48. Densen, P. Challenges and opportunities facing medical education. *Trans. Am. Clin. Climatol. Assoc.* **122**, 48–58 (2011).
49. Craig, L. Service improvement in health care: a literature review. *British Journal of Nursing* vol. 27 893–896 (2018).
50. Hosny, A. & Hugo J W. Artificial intelligence for global health. *Science* vol. 366 955–956 (2019).
51. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat Biomed Eng* **2**, 719–731 (2018).
52. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
53. Gagliardi, G. *et al.* Radiation Dose–Volume Effects in the Heart. *International Journal of Radiation Oncology*Biology*Physics* vol. 76 S77–S85 (2010).
54. Darby, S. C. *et al.* Risk of ischemic heart disease in women after radiotherapy for breast cancer. *N. Engl. J. Med.* **368**, 987–998 (2013).
55. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
56. National Lung Screening Trial Research Team *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011).
57. Douglas, P. S. *et al.* PROspective Multicenter Imaging Study for Evaluation of chest pain: rationale and design of the PROMISE trial. *Am. Heart J.* **167**, 796–803.e1 (2014).
58. Hoffmann, U. *et al.* Design of the Rule Out Myocardial Ischemia/Infarction Using Computer Assisted Tomography: a multicenter randomized comparative effectiveness trial of cardiac computed tomography versus alternative triage strategies in patients with acute chest pain in the emergency department. *Am. Heart J.* **163**, 330–8, 338.e1 (2012).
59. Hoffmann, U., Massaro, J. M., Fox, C. S., Manders, E. & O'Donnell, C. J. Defining normal distributions of coronary artery calcium in women and men (from the Framingham Heart Study). *Am. J. Cardiol.* **102**, 1136–41, 1141.e1 (2008).
60. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care. *Circulation* vol. 117 743–753 (2008).
61. Huang, Y.-L. *et al.* Reliable categorisation of visual scoring of coronary artery calcification on low-dose CT for lung cancer screening: validation with the standard Agatston score. *Eur. Radiol.* **23**, 1226–1233 (2013).
62. Budoff, M. J. *et al.* Coronary artery and thoracic calcium on noncontrast thoracic CT scans: comparison of ungated and gated examinations in patients from the COPD Gene cohort. *J. Cardiovasc. Comput. Tomogr.* **5**, 113–118 (2011).

63. Hecht, H. S. Coronary Artery Calcium Analysis and Reporting on Noncontrast Chest CT Scans: a Paradigm Shift in Prevention. *Current Cardiovascular Imaging Reports* vol. 9 (2016).
64. Lu, M. T. *et al.* Lung Cancer Screening Eligibility in the Community: Cardiovascular Risk Factors, Coronary Artery Calcification, and Cardiovascular Events. *Circulation* **134**, 897–899 (2016).
65. Yang, D. *et al.* Automatic Liver Segmentation Using an Adversarial Image-to-Image Network. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017* 507–515 (2017) doi:10.1007/978-3-319-66179-7_58.
66. Tsao, C. W. & Vasan, R. S. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *Int. J. Epidemiol.* **44**, 1800–1813 (2015).
67. Douglas, P. S. *et al.* Outcomes of anatomical versus functional testing for coronary artery disease. *N. Engl. J. Med.* **372**, 1291–1300 (2015).
68. Hoffmann, U. *et al.* Coronary CT angiography versus standard evaluation in acute chest pain. *N. Engl. J. Med.* **367**, 299–308 (2012).
69. Detrano, R. *et al.* Coronary Calcium as a Predictor of Coronary Events in Four Racial or Ethnic Groups. *New England Journal of Medicine* vol. 358 1336–1345 (2008).

Summary

Deep learning for medical applications has made huge progress in recent years, benefitting from technical advances of computational systems and the availability of large clinical data sets. It has shown great performance, matching and even outperforming human abilities in several areas. Deep learning is especially well suited for image processing tasks, making it ideal for medical imaging applications, e.g. in radiology or radiation oncology. After great successes of deep learning in early proof of concept studies, the research has shifted towards real world applications. In the near future, deep learning is believed to be able to accelerate labor-intensive manual tasks, assist and guide human experts in decision making tasks, or monitor human work and provide quality control in high risk tasks.

The focus of the first study was to develop a deep learning system to automatically segment coronary calcium in chest CT scans and calculate a cardiac risk score. Assessments in large and diverse test cohorts from distinct clinical trials showed that the automatically calculated risk score was a strong predictor for future cardiac events and matched human derived scores. The large test set size of over 20,000 samples showed the great generalizability and robust performance of the presented system and with a processing time of under two seconds per scan, highlighted the high throughput of the system, utilizing state of the art technology.

Based on the experience of the first study, a high resolution heart segmentation deep learning system was developed. The focus of this study was to implement a robust and fast system that is able to segment the heart in cardiac ECG-gated and non-gated CTs to provide a research tool for large clinical studies. The presented system proved its performance and applicability in a series of studies. Two studies successfully used different development versions of this system to assess the associations of whole heart volume and epicardial adipose tissue with future cardiac events in almost 4,000 patients of the PROMISE trial.

The last study showed that a deep learning system developed in cardiovascular radiology can be applied in radiation oncology to automatically segment the heart in radiotherapy planning CTs for breast cancer patients. The system's performance was first tested in a research setting, supporting dosimetrists in their segmentation tasks, where it was able to significantly reduce segmentation time and increase inter-reader variability, with constant segmentation accuracy. In a consecutive test in real world data including over 5,500 clinically accepted and used planning CTs, the deep learning system showed high concordance with manual segmentations with significantly lower failure rate.

Finally, the thesis presents a future perspective of deep learning in general and also of further improvements and applications of the presented methods.

Societal impact and valorizations

In this thesis, state of the art deep learning methods for medical applications were thoroughly described. The presented fully automatic CAC score estimation not only has the potential to support radiologists in their work to increase treatment efficiency and performance but could also be applied to nearly every chest CT scan taken, even if cardiac risk assessment is not the primary reason for taking the scan. For example, this system could be used to alert radiologists scanning individuals for lung cancer assessment that their patient has an increased cardiac risk and trigger a referral to cardiac specialists for follow up treatment. The presented whole heart segmentation system has shown to increase time efficiency of radiation therapy planning in a clinical workflow while maintaining segmentation accuracy and increasing overall segmentation consistency. Furthermore, these deep learning systems have shown that they are able to fast and reliably process very large cohorts of tens or hundreds of thousands of samples, which human experts would simply be unable to handle due to time constraints.

Although deep learning has shown huge success in several fields, it has yet to prove its applicability for real world applications. In recent years the number of deep learning based publications has increased dramatically but their lack of large and distinctive test cohorts often leaves the question for real world applicability and generalizability open. A further problem of medical publications is that training and test cohorts can not, or only with great efforts, be shared with the public. Additionally, sharing of trained medical models is still rare. In this thesis we focused not only on the development of novel systems but also on their real world applicability and their future impact. Therefore, we tested them in several large, independent and distinctive clinical cohorts. Furthermore, we made our full code and trained models publicly available to enable other research groups not only to replicate our results but also to test and apply our methods on data from research groups all over the world, which will further assess their applicability and hence, further enhance research and medical treatment in this field.

To share our code and the trained deep learning models we created dedicated project pages on the lab webpage at *https://aim.hms.harvard.edu*. The code is hosted and maintained at the open-source development platform *www.GitHub.com*. With our open source contributions we aim to have a positive impact on cardiac research and medical treatment. Coronary heart disease is still the most common cause of death in the western civilization. Early cardiac risk prediction has shown to be able to prevent future cardiac events by suggesting life-style changes. The fast automatic risk prediction makes it possible to process every recorded chest CT and assess the cardiac risk of the scanned individual. This may significantly increase the number of individuals with cardiac risk assessment and help prevent future cardiac events.

Although coronary artery calcium represents the current Gold standard for cardiac risk prediction, further advancements in cardiac risk prediction are still desired.

The heart size as well as the amount of fat within the heart are two measures with the potential to further increase the performance of cardiac risk prediction. A fully automatic implementation and application on every recorded chest CT can decrease the number of future cardiac events.

Acknowledgement

I dedicate this thesis to my wife and love Oana. Thank you for your endless support, I wouldn't have been able to do this without you! I want to thank my PhD advisor, supervisor and mentor Dr. Hugo Aerts as well as my PhD supervisor Dr. Udo Hoffmann for their help and support. I also want to thank the assessment committee for their time and effort reviewing this thesis. Furthermore I want to thank all my colleagues and friends I had the opportunity and pleasure to work with on the amazing projects of this thesis. Thank you for all the support and your hard work. And a special thanks goes to my family, my parents and my sister. Finally, I want to NOT thank Covid. I would have done fine without you!

Curriculum vitae

Roman Zeleznik was born in Graz, Austria. He decided early to focus his education onto technical areas and earned his higher school certificate at the secondary college for industrial management in Weiz, Austria. Thereafter, he obtained his Bachelor of Science degree for Information and Computer Engineering at the Graz University of Technology with a focus on network security and data encryption. He enrolled in the Masters program of Information and Computer Engineering, where he focused on Computer Vision and Computational Intelligence. During this time, Roman was awarded with the prestigious Austrian Marshall Plan scholarship that enabled him to conduct his Master's Thesis project at the Massachusetts Institute of Technology in Boston.

Mr. Zeleznik started working in the private sector during his studies where he gained valuable practical experience. After several years he decided to move his focus back to research and started to work at the Institute of Computer Graphics and Vision at the Graz University of Technology. He then had the opportunity to join the lab of Dr. Hugo J.W.L. Aerts in Boston at the Dana-Farber Cancer Institute and started the PhD program at the Maastricht University. During this time he collaborated with several research groups at the MGH, Brigham and Women's Hospital Dana-Farber Cancer Institute and more.

He currently works as a research fellow at Dana-Farber Cancer Institute, Brigham and Women's hospital and Harvard medical school in Boston, Massachusetts, USA.

Scientific publications

Roman Zeleznik, Borek Foldyna, Parastou Eslami, Jakob Weiss, Ivanov Alexander, Jana Taron, Chintan Parmar, Raza M. Alvi, Dahlia Banerji, Mio Uno, Yasuka Kikuchi, Julia Karady, Lili Zhang, Jan-Erik Scholtz, Thomas Mayrhofer, Asya Lyass, Taylor F. Mahoney, Joseph M. Massaro, Ramachandran S. Vasan, Pamela S. Douglas, Udo Hoffmann, Michael T. Lu, Hugo J. W. L. Aerts. 2021. "Deep convolutional neural networks to predict cardiovascular risk from computed tomography." Nature Communications, 12(1), pp.1-9.

Roman Zeleznik, Borek Foldyna, MD; Jakob Weiss, Parastou Eslami, Dennis Bontempi, Michael T. Lu, Udo Hoffmann, Hugo J.W.L. Aerts. 2021. "Deep Learning for fully automatic high resolution heart segmentation in computed tomography scans." Submitted at Elsevier - Medical Image Analysis

Roman Zeleznik*, Jakob Weiss*, Jana Taron, Christian Guthier, Danielle S. Bitterman, Cindy Hancox, Benjamin H. Kann, Daniel W. Kim, Rinaa Sujata Punglia, Jeremy Bredfeldt, Borek Foldyna, Parastou Eslami, Michael T. Lu, Udo Hoffmann, Raymond Mak, Hugo J.W.L. Aerts. 2021. "Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer." npj Digital Medicine

Borek Foldyna, **Roman Zeleznik**, Parastou Eslami, Thomas Mayrhofer, Maros Ferencik, Daniel O Bittner, Nandini M Meyersohn, Stefan B. Puchner, Hamed Emami, Hugo JWL Aerts, Pamela S Douglas, Michael T. Lu, Udo Hoffmann. 2020. "Epicardial Adipose Tissue in Patients With Stable Chest Pain: Insights From the PROMISE Trial." Cardiovascular Imaging, 13(10), pp.2273-2275.

Borek Foldyna, **Roman Zeleznik**, Parastou Eslami, Thomas Mayrhofer, Jan-Erik Scholtz, Maros Ferencik, Daniel O Bittner, Nandini M Meyersohn, Stefan B. Puchner, Hamed Emami, Patricia A. Pellikka, Hugo JWL Aerts, Pamela S Douglas, Michael T. Lu, Udo Hoffmann. 2021. "Small whole heart volume predicts cardiovascular events in patients with stable chest pain: insights from the PROMISE trial." European radiology, pp.1-11.

Yiwen Xu, Ahmed Hosny, **Roman Zeleznik**, Chintan Parmar, Thibaud Coroller, Idalid Franco, Raymond H. Mak, Hugo J.W.L. Aerts. 2019. "Deep learning predicts lung cancer treatment response from serial medical imaging." Clinical Cancer Research, 25(11), pp.3266-3275.

Parastou Eslami, Chintan Parmar, Borek Foldyna, Jan-Erik Scholtz, Alexander Ivanov, **Roman Zeleznik**, Michael T. Lu, Maros Ferencik, Ramachandran S. Vasan, Kristin Baltrusaitis, Joseph M. Massaro, Ralph B. D'Agostino, Thomas Mayrhofer, Christopher J. O'Donnell, Hugo J. W.

L. Aerts, Udo Hoffmann. 2020. "Radiomics of Coronary Artery Calcium in the Framingham Heart Study." Radiology: Cardiothoracic Imaging, 2(1), p.e190119.

Sophia C. Kamran, Thibaud Coroller, Nastaran Milani, Vishesh Agrawal, Elizabeth H. Baldini, Aileen B. Chen, Bruce E. Johnson, David Kozono, Idalid Franco, Nitish Chopra, **Roman Zeleznik,** Hugo J. W. L. Aerts and Raymond Mak. 2020. "The impact of quantitative CT-based tumor volumetric features on the outcomes of patients with limited stage small cell lung cancer." Radiation Oncology, 15(1), pp.1-10.

Ahmed Hosny, Chintan Parmar, Thibaud P. Coroller, Patrick Grossmann, **Roman Zeleznik**, Avnish Kumar, Johan Bussink, Robert J. Gillies, Raymond H. Mak, Hugo J. W. L. Aerts. 2018. "Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study." PLoS medicine, 15(11), p.e1002711.