Moving from traditional methods towards artificial intelligence in cardiovascular research with regular care data

Siegersma, Klaske Rynke

2022

document version
Publisher's PDF, also known as Version of record

Moving from traditional methods towards artificial intelligence in cardiovascular research with regular care data

Klaske Rynke Siegersma

Moving from traditional methods towards artificial intelligence in cardiovascular research with regular care data
Thesis, VU University, Amsterdam, the Netherlands

Cover design: Tynke Siegersma, Klaske Siegersma
Layout: Klaske Siegersma
Printing: ProefschriftMaken.nl
ISBN: 978-94-6423-956-0

The research described in this thesis was supported by a grant of the Dutch Heart Foundation (2018B017 CVON-AI).

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Additional financial support by Chipsoft, Vrije Universiteit and Cardiologie Centra Nederland is gratefully acknowledged.

VRIJE UNIVERSITEIT

Moving from traditional methods towards artificial intelligence in cardiovascular research with regular care data

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op woensdag 5 oktober 2022 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Klaske Rynke Siegersma

geboren te Utrecht

promotoren:	prof.dr. H.M. den Ruijter prof.dr. L. Hofstra
copromotoren:	dr. N.C. Onland-Moret dr. Y. Appelman
promotiecommissie:	prof.dr. J.J. Piek prof.dr. M.L. Bots prof.dr. I. Isgum prof.dr. R. Verheij dr. M.J.M. Cramer dr. J.W.H. Verjans

TABLE OF CONTENTS

Chapter 1

General Introduction

The worldwide prevalence of cardiovascular disease (CVD) has doubled in the past three decades from 271 million to 523 million individuals. Although CVD is still the leading cause of mortality worldwide in men and women, life-expectancy for individuals diagnosed with CVD has steadily increased in high-income countries.[1–3] This can be addressed to targeted reduction in modifiable risk factors, for example smoking and alcohol[4], but also due to improved interventions and new medical therapies.[5] Consequently, the burden of disability due to CVD has increased, which gives a pressing need to focus on efficient interventions and policies, implemented in effective CVD guidelines. Also early recognition and primary prevention using risk-reducing strategies play an important role to decrease the burden of CVD disability. Nonetheless, research into the effectiveness of guidelines and clinical implementation is limited. Furthermore, randomized controlled trials (RCT) are still underpowered for specific subgroups of individuals with CVD. Especially the field of sex-specific knowledge on CVD has long been underexposed, as CVD has historically been seen as a disease for the male half of the world population, although women make up half of all CVD patients.[2] Early recognition and risk-reducing strategies impose a prominent role upon the first line of treatment, e.g. general practitioners or dedicated cardiovascular screening centers.[6]

As such a large part of the world population is suffering from CVD, regular care data is abundantly present in this domain. Regular care data encompasses all the information from a care trajectory of a patient registered by clinicians, but can also include imaging, insurance and pharmacy data. It is therefore a perfect fit for research into the effectiveness of clinical guidelines and the focused studying of specific subgroups of individuals with CVD, as these subgroups are currently underrepresented or underpowered in most clinical trials.[7–9] Especially with the introduction of the electronic health record (EHR) a wide array of medical data has become digitally and directly available. EHRs typically contain all longitudinal information about a patient's care trajectory that is registered at a specific location of care; medical history, diagnostics, therapeutics and interventions.[10] However, EHRs are currently primarily used for the patient's care, whereas it can also be an ideal source of data for clinical research to complement results of RCTs. Research with regular care data can thus help to overcome the knowledge gap as it includes data on all patients, regardless of their sex, age, socio-economic status or ethnicity.[11]

The use of regular care data raises certain issues and asks for new research methodologies to account for these issues. Given the abundance of, potentially unstructured, data, there is a potential use case for artificial intelligence (AI) methods to process and analyse these data. However, as this is a relatively new field of research methodology in healthcare, its clinical value must be proven before implementation in clinical practice. This thesis focuses on the use of research methods ranging from traditional statistics towards AI to generate evidence from regular care data for pressing topics in CVD.

Regular care data: A tool for enrichment of evidence-based medicine

The principles of evidence-based medicine are the backbone of our healthcare practice.[12] These principles imply that decision-making in healthcare is based on a combination of the best available external clinical evidence, a physician's clinical expertise and the patient's wishes and preferences.[13] RCTs are considered as the gold standard to fuel evidence-based medicine in hierarchical mappings of evidence.[14–17] In the past decades efforts were made to gather and systematically combine all available evidence of clinical trials in healthcare to improve patient care.[18]

However, RCTs have several limitations. First, they are expensive[11], due to costs of the experimental clinical procedures, study support staff and site-monitoring costs.[19] Second, this type of trials usually handle strict in- and exclusion criteria, which leads to underrepresentation of multiple subgroups, e.g. women, specifically in the reproductive age[7], the elderly, non-white ethnicities, and patients with comorbidities.[8,9,20] This selective patient inclusion in clinical trials hampers generalisability of results to the general population and application of guidelines in clinical practice.[13] Third, it was shown that patients included in clinical trials on average live longer and more healthy, which is called the 'healthy volunteer inclusion bias'. This makes results not generalisable to the actual population.[21,22] Fourth, RCTs often use composite endpoints to increase the number of events and to reduce follow-up time and trial costs.[23] However, this impedes the interpretation of study results, as components of composite endpoints differ in severity and importance for the individual patient. On top, treatment effect on the different endpoints separately is underreported.[24] Also the use of intermediate endpoints can impede the analysis of patient-centered outcomes. Nevertheless, studying intermediate endpoints increases knowledge on possible causal pathways.[25,26]

Regular care data might complement the results found in RCTs and thereby overcome some of the aforementioned disadvantages that are inherent to RCTs. The greatest potential of regular care data is the opportunity to use already existing and present data to generate new evidence for different research questions, that encompass the uptake and implementation of new guidelines in clinical practice[27], but also study diagnostics, therapies and patient outcomes[28–30], or perform population health research[28]. First, the amount of data and patients in a regular care database usually exceed the number of inclusions in clinical trials.[28] As inclusions in trials are usually limited by costs, the use of regular care data results in a greater sample size at lower costs, which makes analysis in subgroups more feasible. Together with the fact that a broader selection of patients can be included, the problem of underrepresentation of certain subpopulations can be overcome, i.e. in women, as integration of sex in study design is still non-standard practice[31]. Hence, regular care data represents an inclusive patient population and can therefore solve the presence of sex bias in cardiovascular research and results. Second,

a large population also makes research results more generalisable to the contemporary clinical setting. Individuals present in regular care data are an actual representation of the patients that utilise our healthcare resources.[32] Third, regular care data can be linked to other registry databases, which gives insightful information in long-term outcomes and provides the researcher with long-term and high-quality follow-up.[11] Fourth, the use of regular care data removes the need of laborious and time-consuming data collection for research purposes, reducing the workload of researchers and clinicians.

Regular care data is an upcoming, but still underused source of valuable information for research, even though it is only a mouse-click away. Nonetheless, medical data should be structured and cleaned, and preferably complete, to use for traditional research. This is rarely the case with regular care data, as free text fields are abundant, and diagnostics, therapeutics and interventions are driven by medical need. On top, unstructured medical data, e.g. imaging, electrophysiological signals and measurements, are unfit to be directly used in traditional research without accurate interpretation by a healthcare professional. Regular care data thus asks for new research methods.

Moving from traditional research towards artificial intelligence

Regular care data includes many different sources, including, but not limited to: imaging data, lab values, free text of consults, diagnosis, physiological data and genetic data.[10] Traditional statistical methods are not always suitable to process this type of multi-factorial data. On the contrary, AI and its corresponding methods are able to handle large datasets with many different types of data. This makes regular care data especially useful for AI applications. It gives the opportunity to identify patterns in data that have not been described or identified before and that are beyond what humans are able to grasp.

AI can be defined in many ways. Marvin Minsky, the founder of MIT Artificial Intelligence Laboratory, has described AI as a science that makes computers 'go beyond arithmetic' and the ability of those computers to 'imitate the information processes that happen inside human minds'. He outlined this in his 1986 essay entitled 'Why humans think computers can't'.[33] It means that a large amount of data is required to be fed into the computer system to enable computers to identify patterns and imitate information processes that humans acquire.[34] Consequently, the application of AI methods, i.e. machine learning (ML) and deep learning (DL), has recently boosted in healthcare research[35] and makes use of all types of medical data sources available in the EHR including free text fields.

Nonetheless, also the use of AI has certain disadvantages and the use of AI is not always justified. First, critics in healthcare argue that AI is a 'black box' and that it is unknown how the AI makes any decisions[36]. Second, data quality is an important issue. A famous saying used in computer science is 'garbage in, garbage out'. This also applies to AI, meaning that AI is not a panacea that turns unstructured 'rubbish' data into clear

answers to research questions. Third, AI requires appropriate training data. This requires manual labelling[37], making it prone to mistakes and bias, that might seep through in healthcare decision made by AI[38,39]. Fourth, although AI has shown promising results in research settings, it has not yet found its way into daily clinical practice[40], due to a lack of studies that perform external validation and report clinical efficacy[41]. The opportunity of AI to use unstructured data or data with many variables might outweigh the presented disadvantages of AI. However, the use of AI over traditional research methods should be justified by the research question and used data sources to live up to its promises.

Thesis overview

The aim of this thesis was to investigate and explore the use of different traditional statistical methods and AI on regular care data. This was done in order to, firstly, provide insight into pressing topics in CVD, specifically focusing on sex differences and, second, to identify the opportunities for AI in cardiovascular research with regular care data. **Chapter 2** describes the databases of the Cardiology Centers of the Netherlands (CCN). This is a regular care database that contains EHR data on 109,151 individuals with cardiovascular symptoms that are referred by the general practitioner to one of the clinics of the CCN; a cardiac diagnostic screening center that is positioned between the general practitioner's office and the hospital. In the first part of this thesis, we use traditional statistical methods to evaluate the current shift in cardiac practice to replace traditional ECG (electrocardiographic) stress testing by cardiac imaging for the diagnosis of coronary heart disease (CHD), a subtype of CVD that accounts for almost 50% of all CVD deaths worldwide.[2] CHD evolves from atherosclerosis of the coronary vasculature towards build-up of plaques and can finally lead to possibly lethal occlusion of the coronaries, resulting in myocardial infarction. Regular practice for detection of CHD has been ECG stress testing, either by bike or treadmill. Two RCTs, comparing usual care that included ECG stress testing with cardiac computed tomographic (CT) imaging, both coronary calciumscore[42] and coronary CT angiography, for diagnosis of CHD, showed improved results in patients with chest pain that underwent a so-called CT-first strategy.[43,44] Consequently, international guidelines have taken on a more prominent role for cardiac CT and ECG stress testing is gradually phased out. EHR data gives the opportunity to evaluate the effect of this shift of diagnostics for CHD in an actual regular care population. In **Chapter 3** we evaluate the effect of a CT-first strategy on mortality risk in individuals that present with chest pain at one of the CCN centers. In **Chapter 4**, the results of the CT-scan were used to evaluate their prognostic value in women and men separately. **Chapter 5** utilizes a sex-specific approach to evaluate the association of self-reported New York Heart Association (NYHA)-class, a subjective measure of functional performance and condition, and mortality (**Chapter 5A**). **Chapter 5B** dives deeper into sex differences and studies the sex-specific

mediating effect of the ECG stress test on the relation between NYHA-class and mortality in order to evaluate its value in clinical practice.
In the second part of this thesis, different cardiac diagnostic modalities are used to show the application and value of AI in clinical research. First, a pipeline is described in which widespread word-embedding models and string matching are used to identify medications and adverse drug reactions in consult texts (**Chapter 6**). This is followed by a narrative overview of the application of AI in cardiac imaging in **Chapter 7**. In **Chapter 8** EHR data from the CCN database is used to evaluate the sex-stratified performance of the pretest probability for CHD, which is used for referral to cardiac CT. AI is used to see whether sex-specific models with more data improve this referral. **Chapter 9** shows the results of a study that uses normal ECGs to classify sex. This chapter emphasises the necessity to perform sex-stratified research by revealing ECG features that are associated with mortality, which have not been previously identified. The concluding chapter (**Chapter 10**) summarises the results of this thesis and puts them in the perspective of contemporary cardiac practice and research. It provides an outlook into the development and uptake of cardiac AI and the role of traditional statistics and epidemiology in this development.

REFERENCES

1. Taylor CJ, Ordóñez-Mena JM, Roalfe AK, et al. Trends in survival after a diagnosis of heart failure in the United Kingdom 2000-2017: population based cohort study. *BMJ*. 2019;364:1-10. doi:10.1136/bmj.l223
2. Roth GA, Mensah GA, Johnson CO, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol*. 2020;76(25):2982-3021. doi:10.1016/j.jacc.2020.11.010
2. de Boer, A.R.; van Dis, I.; Wimmers, R.H.; Vaartjes, I.; Bots ML. Cijfers- Hart En Vaatziekten in Nederland 2020. *Cent Bur voor Stat en Niv*. Published online 2020.
3. Abbafati C, Abbas KM, Abbasi-Kangevari M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10258):1223-1249. doi:10.1016/S0140-6736(20)30752-2
4. IOM (Institute of Medicine). Reducing the Burden of Cardiovascular Disease: Intervention Approaches. In: Fuster V, Kelly BB, eds. *Promoting Cardiovascular Health in the Developing World: A Challenge to a Acheive Global Health*. The National Academies Press; 2010:185-274.
5. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice Developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of. *Eur Heart J*. Published online 2021:1-111. doi:10.1093/eurheartj/ehab484
6. Pilote L, Raparelli V. Participation of Women in Clinical Trials: Not Yet Time to Rest on Our Laurels. *J Am Coll Cardiol*. 2018;71(18):1970-1972. doi:10.1016/j.jacc.2018.02.069
7. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA Intern Med*. 2014;174(11):1868-1870. doi:10.1001/jamainternmed.2014.4758
8. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals. *J Am Med Assoc*. 2007;297(11):1233-1240. doi:10.1001/jama.298.1.39-b
9. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405. doi:10.1038/nrg3208
10. Mc Cord KA, Al-Shahi Salman R, Treweek S, et al. Routinely collected data for randomized trials: Promises, barriers, and implications. *Trials*. 2018;19(1):1-9. doi:10.1186/s13063-017-2394-5
11. Sackett DL. Evidence-based medicine. Semin Perinatol. 1997;21(1):3-5. doi:10.1016/S0146-0005(97)80013-4
12. Greenhalgh T, Howick J, Maskrey N, et al. Evidence based medicine: A movement in crisis? *BMJ*. 2014;348(June):1-7. doi:10.1136/bmj.g3725
13. Canadian Taks Force on the Periodic Health Examination. Periodic health examination. *Can Med Assoc J*. 1979;121(9):1193-1254. doi:10.7326/0003-4819-62-4-853
14. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest*. 1989;95(2):1S-4S. doi:10.1378/chest.105.2.647b
15. Bob Phillips, Ball C, Sackett D, et al. Oxford Centre for Evidence-Based Medicine: Levels of Evidence (March 2009). Centre for Evidence-Based Medicine. Published 2011. Accessed June 21, 2021. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009
16. Burns PB, Rohrich RJ, Chung KC. The

levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg*. 2011;128(1):305-310. doi:10.1097/PRS.0b013e318219c171
17. Levin A. The Cochrane Collaboration. *Ann Intern Med*. 2001;135(4):309-312. doi:10.1177/0193945913491839
18. Sertkaya A, Wong HH, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin Trials*. 2016;13(2):117-126. doi:10.1177/1740774515625964
19. Van der Marck MA, Melis RJF, Rikkert MGMO. On evidence-based medicine. *Lancet*. 2017;390(10109):2244-2245. doi:10.1016/S0140-6736(17)32851-9
20. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am J Epidemiol*. 2007;165(8):874-881. doi:10.1093/aje/kwk075
21. Leening MJG, Heeringa J, Deckers JW, et al. Healthy volunteer effect and cardiovascular risk. *Epidemiology*. 2014;25(3):470-471. doi:10.1097/EDE.0000000000000091
22. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite Outcomes in Randomized Trials: Greater Precision but with Greater Uncertainty? *J Am Med Assoc*. 2003;289(19):2554-2559. doi:10.1001/jama.289.19.2554
23. Ferreira-González I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: Systematic review of randomised controlled trials. *BMJ*. 2007;334(7597):786-788. doi:10.1136/bmj.39136.682083.AE
24. Lonn E. The use of surrogate endpoints in clinical trials: Focus on clinical trials in cardiovascular diseases. *Pharmacoepidemiol Drug Saf*. 2001;10(6):497-508. doi:10.1002/pds.654
25. Asmar R, Hosseini H. Endpoints in clinical trials: Does evidence only originate from hard' or mortality endpoints? *J Hypertens*. 2009;27(SUPPL. 2):45-50. doi:10.1097/01.hjh.0000354521.75074.67
26. Tong ST, Sabo RT, Hochheimer CJ, et al. Uptake of statin guidelines to prevent and treat cardiovascular disease. *J Am Board Fam Med*. 2021;34(1):113-122. doi:10.3122/jabfm.2021.01.200292
27. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health*. 2016;37:61-81. doi:10.1146/annurev-publhealth-032315-021353
28. Shah SM, Khan RA. Secondary use of electronic health record: Opportunities and challenges. *IEEE Access*. 2020;8(July):136947-136965. doi:10.1109/ACCESS.2020.3011099
29. Boulton C, Wilkinson JM. Use of public datasets in the examination of multimorbidity: Opportunities and challenges. *Mech Ageing Dev*. 2020;190(xxxx):111310. doi:10.1016/j.mad.2020.111310
30. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. *Nature*. 2019;575(7781):137-146. doi:10.1038/s41586-019-1657-6
31. Whittaker H, Quint JK. Using routine health data for research: The devil is in the detail. *Thorax*. 2020;75(9):714-715. doi:10.1136/thoraxjnl-2020-214821
32. Minsky M. Why People Think Computers Can't. *AI Mag*. 1982;3(4):3-15.
33. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Heal Inf Sci Syst*. 2014;2(3):1-10. doi:10.1145/2347736.2347741
34. Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering*. 2020;6(3):291-301. doi:10.1016/j.eng.2019.08.015
35. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black Box" Medi-

cine? *Ann Intern Med*. 2020;172(1):59-61. doi:10.7326/M19-2548
36. Sermesant M, Delingette H, Cochet H, Jaïs P, Ayache N. Applications of artificial intelligence in cardiovascular imaging. *Nat Rev Cardiol*. 2021;0123456789. doi:10.1038/s41569-021-00527-2
37. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res*. 2018;20(5). doi:10.2196/JMIR.9134
38. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digit Med*. 2020;81. doi:10.1038/s41746-020-0288-5
39. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *npj Digit Med*. 2018;1(1):10-12. doi:10.1038/s41746-018-0048-y
40. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*. 2021;31(6):3797-3804. doi:10.1007/s00330-021-07892-z
41. Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte M, Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. *J Am Coll Cardiol*. 1990;15(4):827-832. doi:10.1016/0735-1097(90)90282-T
42. The SCOT-HEART Investigators. Coronary CT Angiography and 5-Year Risk of Myocardial Infarction. *N Engl J Med*. 2018;379(10):924-933. doi:10.1056/nejmoa1805971
43. Douglas PS, Hoffmann U, Patel MR, et al. Outcomes of anatomical versus functional testing for coronary artery disease. *N Engl J Med*. 2015;372(14):1291-1300. doi:10.1056/NEJMoa1415516

Chapter 2

Routine clinical care data from thirteen cardiac outpatient clinics: Design of the cardiology centers of the Netherlands (CCN) database

Klaske R. Siegersma[¥], Sophie H. Bots[¥]. N. Charlotte Onland-Moret, Folkert W. Asselbergs, G. Aernout Somsen , Igor I. Tulevski, Hester M. den Ruijter[*], Leonard Hofstra[*]

[¥,*] These authors contributed equally

BMC Cardiovascular Disorders 2021; 21(1): 1-9

ABSTRACT

Background Despite the increasing availability of clinical data due to the digitalisation of healthcare systems, data often remain inaccessible due to the diversity of data collection systems. In the Netherlands, Cardiology Centers of the Netherlands (CCN) introduced "one-stop shop" diagnostic clinics for patients suspected of cardiac disease by their general practitioner. All CCN clinics use the same data collection system and standardised protocol, creating a large regular care database. This database can be used to describe referral practices, evaluate risk factors for cardiovascular disease (CVD) in important patient subgroups, and develop prediction models for use in daily care.

Construction and Content The current database contains data on all patients who underwent a cardiac workup in one of the 13 CCN clinics between 2007 and February 2018 (n=109,151, 51.9% women). Data were pseudonymised and contain information on anthropometrics, cardiac symptoms, risk factors, comorbidities, cardiovascular and family history, standard blood laboratory measurements, transthoracic echocardiography, electrocardiography in rest and during exercise, and medication use. Clinical follow-up is based on medical need and consisted of either a repeat visit at CCN (43.8%) or referral for an external procedure in a hospital (16.5%). Passive follow-up via linkage to national mortality registers is available for 95% of the database.

Utility and Discussion The CCN database provides a strong base for research into historically underrepresented patient groups due to the large number of patients and the lack of in- and exclusion criteria. It also enables the development of artificial intelligence-based decision support tools. Its contemporary nature allows for comparison of daily care with the current guidelines and protocols Missing data is an inherent limitation, as the cardiologist could deviate from standardised protocols when clinically indicated.

Conclusion The CCN database offers the opportunity to conduct research in an unique population referred from the general practitioner to the cardiologist for diagnostic workup. This, in combination with its large size, the representation of historically underrepresented patient groups and contemporary nature can expand our knowledge of cardiovascular diseases.

BACKGROUND

Cardiovascular diseases (CVD) remain an important cause of death and disability worldwide.[1,2] The digitalisation of the healthcare system has made a wealth of clinical care data available for researchers.[3–6] This provides a unique opportunity for researchers to evaluate pressing topics in cardiovascular medicine. The added value of clinical care data in cardiovascular research is threefold. First, clinical care data better reflect the current real-world situation in healthcare with regard to clinical presentation of disease and representation of patient groups. This is especially relevant for patient groups that have historically been underrepresented in clinical studies such as women[7], the elderly[8] and patients with multimorbidity[9]. CVD in women may be different from CVD in men in several aspects, including the clinical presentation, the effect of traditional risk factors and presence of female-specific risk factors related to pregnancy and menopause, and the efficacy of treatment.[10] Elderly patients and those with multimorbidity also need to be studied to combat the rising prevalence of CVD risk factors such as hypertension, diabetes mellitus (DM) and obesity.[11,12] Second, clinical care data contain a large number of individuals and wide range of clinical measurements, a combination that is difficult to obtain within a research setting. This facilitates the development of prediction models and decision support tools using artificial intelligence methods that can subsequently Ibe implemented within the healthcare system. These tools can help healthcare professionals to interpret large amounts of patient data and assist healthcare decision-making. Third, researchers can use clinical care data to evaluate the current state of clinical practice, adherence to guidelines and develop treatment and referral strategies that better suit the current presentation of patients suspected of CVD.

However, data from earlier stages in the clinical care pathway remain difficult to access due to the smaller size of single general practitioner (GP) offices and the diversity of data collection systems. To close this gap, a collaboration was set up between the University Medical Center Utrecht (UMCU) and Cardiology Centers of the Netherlands (CCN), an organisation of 13 cardiac outpatient clinics that operate between the GP and the hospital cardiologist. In the Netherlands, CCN introduced "one-stop shop" cardiac outpatient clinics to facilitate efficient diagnostic workup for cardiac disease and fast diagnosis of potential life-threatening pathologies. GPs can refer their patients to a CCN clinic for cardiac workup when they suspect their patient suffers from cardiac disease. All CCN clinics perform the same standardised protocol and store their data in a shared data collection system. Follow-up appointments and results from referrals for advanced cardiac imaging or cardiac interventions are stored in the same system. As a result of this set-up, CCN offers a unique opportunity to obtain semi-structured data on a large group of patients at an early stage of the regular care pathway.

The aim of this paper is to describe the CCN clinical care database. The database contains

data on a large number of individual patients and a wide range of standardised characteristics from a unique population situated between the GP and the hospital cardiologist. The clinical nature of the database ensures that it reflects the patient population currently seen in daily care, including those that may be underrepresented in clinical research. The database can be used to describe current clinical practice, evaluate the prevalence of cardiovascular risk factors and their relation to CVD, and develop prediction algorithms that have the potential to be implemented in daily care.

CONSTRUCTION AND CONTENT

Data generation at CCN clinics

Baseline examination

Every patient referred to one of the CCN clinics underwent a standardised diagnostic workup. This protocol consisted of transthoracic echocardiography (TTE) and ultrasound imaging of the carotid arteries, electrocardiography at rest (ECG) and during exercise (stress ECG), a laboratory test, and a consult with a nurse during which self-reported anthropometrics, symptoms, cardiovascular risk factors and comorbidities were registered. Past medication use and cardiovascular history were also recorded, as well as on-site clinical diagnoses made by the cardiologist. An overview of all stored clinical characteristics can be found in Table 1.

Body mass index was calculated based on self-reported height and weight. Blood pressure was measured with a Microlife WatchBP. TTE was performed with a General Electric Vivid E6 or E7 echocardiography device. Blood samples were analysed with the Roche Reflotron Sprint system. The ECG was recorded with the Welch Allyn Cardioperfect Pro recorder in supine position with 12 leads. The stress ECG was performed on a watt bike from Lode Corival Eccentric with simultaneous blood pressure measurements (Medtronic BL-6 Compact) and ECG recording (Welch Allyn Cardioperfect recorder). Raw data of the ECG, stress ECG and TTE were not available. Medication and diagnoses were recorded as semi-structured text.

While CCN has standardised and uniform diagnostic workup protocols for every patient, in practice a cardiologist may deviate from this protocol when this is clinically indicated. For example, the cardiologist may choose not to perform a stress ECG in patients with a contra-indication to the procedure, such as very high systolic blood pressure.[13] This introduces missing data, illustrated by the baseline stress ECG data which were missing for 25% of patients in the CCN database (Figure 1).

Information collected during a patient's clinical trajectory within CCN

After the first visit, patients may enter a clinical trajectory during which one or more return

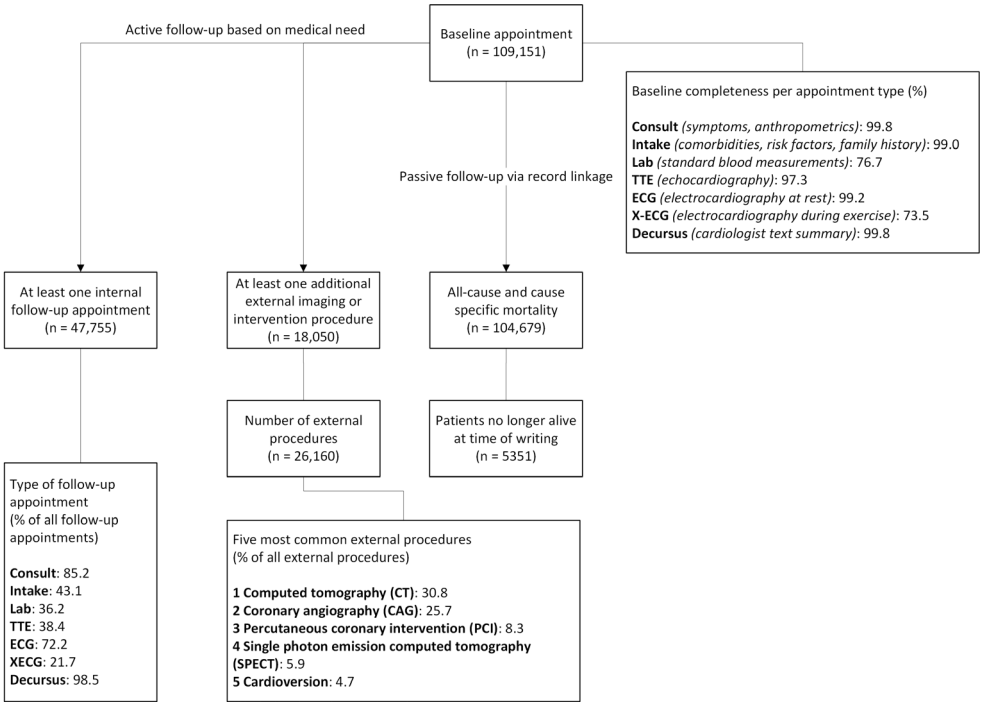


Figure 1 Overview of patient flow and completeness of measurements in the CCN database

visits to a CCN clinic are planned. Information collected during these clinical follow-up visits was also stored in the CCN database. This clinical follow-up was not standardised but rather based on medical need. As a result, clinical follow-up varies across patients in frequency, duration, and measurements obtained. During these clinical follow-up visits either all or some components of the standard screening protocol were repeated, with rest ECG being repeated most frequently (Figure 1).

Patients in need of additional imaging or cardiac interventions based on the result of their initial CCN workup were referred to a nearby hospital as these facilities were not available at the CCN clinics. The referral itself and the summarised text results of these procedures were stored in the CCN database (Table 1). Computed Tomography (CT) scans were performed most often, comprising 30.8% of all external procedures. The five most common external procedures can be found in Figure 1.

Database construction

Data extraction, cleaning and storage

We extracted all data generated by CCN up to February 2018 from their data collection system. These raw files were cleaned and processed using SAS (SAS Institute Inc., North Carolina, USA) to create a relational database. This process included separating first visit

Table 1 Overview of all features stored in the database

Phase	Measurement
Baseline (2007-Feb 2018)	*Consult* - Presence and characteristics of cardiac symptoms (chest pain, dyspnoea, fatigue, palpitations, collapse, heart murmurs) - Anthropometrics (height, weight, hip circumference, blood pressure, heart rate, heart and breathing sounds, pulse, palpation) *Intake* - Behavioural cardiovascular risk factors (smoking, alcohol use) - Comorbidities (diabetes mellitus, hypertension, dyslipidaemia) - Family history of cardiovascular disease (atherosclerosis, sudden death, cardiomyopathy, arrhythmia) *Lab* - Lipids (total, high density, and low density cholesterol, triglycerides) - Potassium, sodium, haemoglobin, glucose - Glomerular filtration rate - Lipoprotein A, brain natriuretic protein, thyroid stimulation hormone *TTE* - M-mode (dimensions of aorta and left heart chambers) - Two-dimensional (evaluation of function and shape of all heart chambers and valves) - Colour Doppler (valve insufficiencies and septum defects) - Spectral Doppler (left ventricular diastolic function and gradients over valves) - Thickness of the intima media (left and right, anterior and posterior) *ECG* - Duration of defined ECG intervals and complexes (RR, PR, QRS, QT) - ST depression, elevation, negative T-top, QRS axis - Dilatation of left and right atrium, intraventricular conduction delay, left ventricular hypertrophy *Stress ECG* - Protocol, device, target heart rate, use of β-blocker before exercise test - ECG characteristics, blood pressure and heart rate before and during exercise test - Duration and load of exercise test, exercise tolerance, reason to stop exercise test - Arrhythmia or angina symptoms during exercise test, left ventricular hypertrophy *Decursus* - Cardiologist summary of visit (free text) *Medication* - Cardiovascular medication use grouped by researchers - Date medication was started and date it was ended when applicable *Diagnosis*; - Cardiovascular diagnosis defined by researchers - Cardiovascular risk factor diagnosis defined by researchers - Date of diagnosis

Phase	Measurement
Follow-up (2007 - Feb 2018)	Consult, Intake, Lab, TTE, ECG, Stress ECG and Decursus as described for baseline *External procedures* - External procedure performed and location where it was performed - External procedure grouped by researchers - Date of appointment
Record linkage (2019)	- All-cause mortality - Educational level - Ethnicity - Personal income - Cause-specific mortality

TTE: transthoracic echocardiogram, ECG: electrocardiogram.

(baseline) data from follow-up visits, filtering out duplicated or empty entries and removing completely empty variables, streamlining variable names, and organising the data by type of clinical measurement (e.g. combine all laboratory measurements in one data table), among others. Raw unstructured text fields were checked for personal information, which was subsequently either removed while keeping the text field intact or the information was recoded into a new variable that no longer contained the personal information.

Raw medication use and diagnosis text data were structured into binary variables using text retrieval methods in R (R Core Team, Vienna, Austria). Medication entries were grouped into 23 categories of relevant cardiovascular medications based on either the brand name or the generic name, depending on which one was available (Supplementary table 1). Diagnoses were divided into (i) cardiovascular disease and (ii) conditions that are risk factors for cardiovascular disease. The first category was subdivided into 5 subgroups, the second one into 4 subgroups (Supplementary table 2).

The raw data and the clean relational database are stored within the UMCU infrastructure. The raw data is not available for researchers due to privacy constraints and is kept by the data manager. The anonymised versions of raw unstructured text fields are available, including raw medication and diagnosis data. Researchers can contact the authors for collaboration and access to the UMCU infrastructure. When the collaboration and the research topic have been agreed upon, external collaborators can get access to both the CCN database and all services and programmes supported by the UMCU. This includes artificial intelligence and advanced statistical programs. All work within the UMCU infrastructure will be stored, including analysis scripts and results. Access to the UMCU infrastructure will be retracted after the project has finished.

Passive and active follow-up outside the clinical trajectory

The CCN database has been linked to the national database of Statistics Netherlands for

passive follow-up for all-cause and cause-specific mortality, and enrichment of the dataset with demographic and socioeconomic data. Linkage was successful for 95.9% of the database (Figure 1). Failure to link likely occurred because a patient moved between their CCN visit and the moment of linking, as postal code was one of the linking factors. Linking of the CCN database with Statistics Netherlands was deemed appropriate by the ethical committee of Statistics Netherlands as it was in line with the CCN project aims.
Access to the following data was requested and granted: (i) all-cause and cause-specific mortality, (ii) education level and personal income and (iii) personal records database, which among others contains information on country of birth. Access to the personal records database also enables researchers to obtain a matched sample of the general population for comparison with the CCN population. In the future, the CCN database will be linked to other registries, such as the national hospitalisation registry, to obtain information on a more diverse set of outcome measures. Patients could not be contacted for additional baseline questionnaires or active follow-up due to the pseudonymised nature of the database.

Missing data

Diagnostic procedures, treatments and follow-up of the patients were performed at the discretion of the treating cardiologist and thus driven by medical indication. This results in missing data for both baseline and follow-up visits. For example, more advanced biomarkers such as brain natriuretic peptide or high-sensitivity troponin will only be measured if the cardiologist suspects serious cardiac problems. Similarly, patients without entries in the medication or diagnosis file can be assumed to not use medication or be free of disease. Imputation strategies can be applied to deal with the missing values, but the preferred strategy depends on whether the data is likely to be missing at random or not. Researchers should be aware of the assumptions they make and describe these in their methods section.

Patient privacy

The CCN data were made available under implied consent and transferred to the UMCU under the Dutch Personal Data Protection Act. Patients were assigned a unique patient number that cannot be traced back to an individual without access to the original CCN data system, which is not available to UMCU researchers. This results in a pseudonymised database. The Medical Research Ethics Committee of the UMCU declared that the Medical Research Involving Human Subjects does not apply to this study. Unstructured text fields containing personal information were anonymised using an anonymisation programme[14] before being included in the final research database.

Table 2 Baseline characteristics of patients in the CCN database

Variables	Whole database	Women	Men	Missing data (%)
n	109,151	56,628	52,524	
General				
Women (n, %)	56,628 (51.9)			
Age (years)	56 (15)	57 (15)	56 (15)	
Age categories (n, %)				
>50	33,165 (30.4)	16,954 (29.9)	16,211 (30.9)	
50-64	41,273 (37.8)	20,859 (36.8)	20,414 (38.9)	
65-74	22,931 (21.0)	12,152 (21.5)	10,779 (20.5)	
75≤	11,781 (10.8)	6662 (11.8)	5119 (9.7)	
Body mass index (kg/m2)	27.4 (20.0)	27.3 (20.2)	27.5 (19.8)	2.9
Systolic blood pressure (mmHg)	141 (22)	140 (23)	143 (20)	2.9
Current smoker (n,%)	40,139 (36.8)	20,712 (36.6)	19,427 (37)	8.9
Ever smoker (n,%)	71,659 (65.7)	35,508 (62.7)	36,151 (68.8)	8.8
Cardiovascular disease (n, %)				
History of CVD[1]	16,311 (14.9)	6483 (11.4)	9828 (18.7)	
Family history of CVD[2]	71,148 (65.2)	39,318 (69.4)	31,830 (60.6)	17.8
History of other cardiovascular conditions[3]	23,957 (21.9)	11,804 (20.8)	12,153 (23.1)	
Comorbidities (n, %)				
Hypertension	32,460 (29.7)	17,290 (30.5)	15,270 (28.9)	2.5
Dyslipidaemia	16,978 (15.6)	8148 (14.4)	8830 (16.8)	2.5
Diabetes mellitus	8709 (8.0)	3967 (7.0)	4742 (9.0)	2.6
Number of comorbidities				
0	64,199 (58.8)	33,799 (59.9)	30,400 (57.9)	
1	28,705 (26.3)	15,081 (26.6)	13,624 (25.9)	
2	11,001 (10.1)	5392 (9.5)	5609 (10.7)	
3	2382 (2.2)	1125 (2.0)	1257 (2.4)	

All values are given as mean (SD) unless otherwise specified. [1]History of CVD: diagnosis of heart failure, coronary heart disease, cerebrovascular disease or congenital heart disease before baseline appointment, or invasive cardiac intervention. [2] Family history of CVD: family history of atherosclerosis, sudden death, cardiomyopathy or arrhythmia. [3] History of the cardiovascular conditions: diagnosis of arrhythmia, valvular disease, cardiomyopathy, atherosclerosis, peripheral artery disease or abdominal aneurysm before baseline appointment, or non-invasive cardiac or peripheral intervention. CVD: cardiovascular disease

Content: Describing the CCN study population

The CCN database contains data from 109,227 patients referred to one of the CCN clinics between February 2007 and February 2018 (Supplementary Figure 1). Patients with missing data on age or sex or without records of their CCN visit were excluded (n=76), bringing the total to 109,151 individuals with a mean age of 56 (±15) years, of which 51.9% were women. About a third of the patients were 65 years or older and 12% had two or more comorbidities.

Patients had a mean body mass index of 27.4 (±20) kg/m^2 and an average systolic blood pres sure of 141 (±22) mmHg. The majority of patients had a positive cardiovascular family history (65.2%) and 14.9% of patients suffered from cardiovascular disease at baseline. Approximately one third of patients were current smokers (36.8%), 29.7% had hypertension, 15.6% had dyslipidaemia and 8% had DM (Table 2).

The majority of patients (56.1%, n=61,232) only had a baseline visit, 17.5% (n=19,111) had one follow-up visit at CCN, and 26.3% (n=28,808) had three or more follow-up visits at CCN. Compared with patients who were seen once, those with at least one clinical follow-up appointment were older at baseline (60 vs 54 years), had a higher systolic blood pressure (145 vs 138 mmHg) and were more often current smokers (41.1% vs 33.4%). In addition, they more often had a history of cardiovascular disease (20.6% vs 10.5%), prevalent cardiovascular risk conditions (30.0% vs 15.5%), and comorbidities (Supplementary table 3). In total, 18,050 (16.5%) patients were referred for an external procedure (Figure 1). Compared with patients who were not referred, patients with at least one external procedure were older at baseline (60 vs 56 years) and had a higher prevalence of comorbidities and CVD history (21.3% vs 13.7%). Women were less often referred for an external procedure (46.1% vs 53.0%) (Supplementary table 4).

The CCN database consists of data derived from medical care and thus participants were not actively recruited, nor were there explicit in- and exclusion criteria. Data on patients who were not referred to CCN are not available, so we were unable to compare patients referred to CCN with those who were not. However, to approximate this comparison, we compared the socioeconomic characteristics of the CCN database to an age- and sex-matched sample of the general population. Patients referred to CCN were more often of Dutch descent (77.2% vs 70.8%) and had a higher median annual personal income (€27,914 vs €22,270) than the general population (Table 3).

UTILITY AND DISCUSSION

Utility: Intented use and database benefits

The main strength of the CCN database lies in its combination of a large study population and a large number of different, and sometimes longitudinal, measurements per individual. Such data is difficult to obtain in cohorts specifically set up for research as

Table 3 Sociodemographic characteristics of the CCN database and a sample of the general population matched on year of birth and sex.

	CCN database	General population
n	104,519[1]	104,519
Origin[2] (n, %)		
Native Dutch	80,692 (77.2)	74,042 (70.8)
First generation immigrant	15,731 (15.1)	24,592 (23.5)
Second generation immigrant	8096 (7.7)	5884 (5.6)
Annual personal income (€)	27,914 [14,822-47,344]	22,270 [11,900-38,758]
Annual personal income groups (n, %)		
Negative or zero	4760 (4.6)	5701 (5.5)
<€20.000	33,209 (31.8)	33,048 (31.6)
€20.000-€50.000	42,325 (40.5)	33,906 (32.4)
€50.000-€100.000	18,204 (17.4)	10,530 (10.1)
€100.000-€200.000	4197 (4.0)	1675 (1.6)
≥€200.000	1121 (1.1)	337 (0.3)
Not available	703 (0.7)	19,321 (18.5)

Values are given as median (interquartile range, IQR) unless otherwise specified. [1]Year of birth could not be re-calculated for 160 study participants, so these could not be matched with the general population and are thus removed from this table. [2] Origin was defined as (i) Native Dutch; both parents born in the Netherlands, (ii) First generation immigrant; person born outside the Netherlands with at least one parent born outside the Netherlands, (iii) Second generation immigrant; person born in the Netherlands with at least one parent born outside the Netherlands.

funds are often not sufficient to cover inclusion of a large population and collection of a large number of (longitudinal) measurements. In addition, the CCN database captures a unique population situated between the GP and the hospital that is rarely seen in clinical studies.

Clinical care databases like the CCN database can make important contributions to three areas of research due to some of their inherent characteristics. First, these databases reflect the population currently seen in clinical care and thus include groups that are traditionally underrepresented in research. We show that women comprise 52% of the CCN database, providing a valuable foundation for research into both differences between the sexes and women-specific cardiovascular disease presentations and risk factors.[15,16] Similarly, the CCN dataset contains 11,781 patients aged 75 years and older and 13,383 patients with two or more comorbidities, offering researchers an opportunity to verify if study outcomes also apply to these patient groups. These numbers illustrate the potential value of the CCN database for addressing research questions about underrepresented patient groups that have remained unanswered due to scarcity of data.

Second, the size of clinical care databases that combine a large study population with a large number of measurements per individual creates opportunities for the application of artificial intelligence methods. The CCN database contains more than 300 informative

features on over 100.000 patients that can be used for the development of artificial intelligence-based prediction algorithms and decision support tools. In addition, the CCN database contains several anonymised Dutch free text fields, which can be used for the development of text analysis algorithms specific for Dutch clinical notes. This is an important area of research, as many existing text analysis resources are based on English clinical text.[17] These programmes can subsequently be used to extract and structure valuable information from free text and turn it into a usable format for researchers.
Third, clinical care databases reflect medical practice allowing for comparisons between clinical care and the recommendations in the prevailing guidelines. Such perspectives spark debate on inconsistencies that may exist between guidelines and current practice. The CCN database functions in this case as a tool to bridge the gap between guidelines based mainly on clinical research and the reality of daily cardiac care.

Discussion: Compare performance and functionality with similar existing databases

However, the CCN database also has some limitations that need to be addressed. We will discuss the two main ones, data quality and generalisability.

Data quality: Missing data and measurement errors

The data within the CCN database was collected for care purposes and not for research. As a result, data collection and follow-up during the medical trajectory are not uniform across patients. Similarly, the database may not contain all clinical information researchers need, such as highly specific biomarkers, because these are not normally collected in daily care. Furthermore, raw ECG data and echocardiographic images were saved to a different system than the standardised clinical data and were thus not stored in the CCN database. These limitations are in part inherent to the database, so researchers should consider whether the CCN database is 'fit for purpose' for their specific research question. However, some of these limitations can be addressed and alleviated. To obtain standardised follow-up for all individuals in the CCN database, we performed record linkage for all-cause and cause-specific mortality. We plan to include follow-up for non-fatal outcomes in the future, as these outcomes are clinically relevant for the relatively young and healthy CCN population. To alleviate the issue of missing data on important confounders such as socioeconomic status, we enriched the CCN database with information on ethnicity, educational level and personal income through record linkage. Text mining approaches can be used to further enrich the CCN database if the required information can be found within the unstructured text fields. Available missing data techniques such as multiple imputation can be used to address remaining missing values as long as researchers carefully consider the assumptions underlying these techniques.
Data collection and entry in the CCN database is not checked as vigorously as in databases created for research, so data entry mistakes and slightly differential measurement

practices across CCN clinics may introduce measurement error and misclassification. We have tried to correct the most obvious data entry errors to reduce its effect, but researchers should consider the possibility of differential measurement error and the resulting risk of misclassification bias when interpreting their results.

Generalisability and comparison to other databases

The CCN database is comprised of patients who were referred by their GP on suspicion of cardiac disease. We were unable to compare those included in the CCN database with those who were not referred, but we were able to approach this comparison by using an age- and sex-matched sample from the general population. We show that CCN patients have a higher socio-economic status and are more often native Dutch compared with the general population. Moreover, the prevalence of DM in the CCN database seems to be similar to that in the Netherlands as a whole[18], while we expected a higher prevalence given that CCN screens patients at elevated cardiovascular disease risk. However, GPs may refer DM patients with cardiac complaints to a DM-specific outpatient clinic instead of a CCN clinic, resulting in a low DM prevalence within the CCN database. This suggests there is some selection bias occurring within the clinical care pathway, where relatively healthy Dutch patients with higher socio-economic status are more often referred to a CCN clinic than those with lower socio-economic status or those of non-Dutch descent.

There are examples of other clinical care databases such as the hospital-based UPOD database[3] and the Julius General Practitioner's Network[19]. However, these include distinctively different patient populations, as the first collects data from within the hospital and the second from within GP practice. The CCN database is unique in that it captures the patients in between these two.

CONCLUSION

The CCN database is a regular care database containing data from 109.151 patients collected between 2007 and 2018. This database offers the opportunity to perform research in a unique study population that reflects the patient population seen in daily cardiology practice, including women, the elderly, and patients with multiple comorbidities. The size of this database facilitates the application of artificial intelligence methods. Moreover, the features in the database make it possible to describe current cardiology practice and evaluate this against guidelines based primarily on results from clinical trials.

REFERENCES

1. Vos T, Abajobir AA, Abbafati C, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet*. 2017;390(10100):1211-1259. doi:10.1016/S0140-6736(17)32154-2
2. Roth GA, Abate D, Abate KH, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2018;392(10159):1736-1788. doi:10.1016/S0140-6736(18)32203-7
3. Ten Berg MJ, Huisman A, Van Den Bemt PMLA, Schobben AFAM, Egberts ACG, Van Solinge WW. Linking laboratory and medication data: New opportunities for pharmacoepidemiological research. *Clin Chem Lab Med*. 2007;45(1):13-19. doi:10.1515/CCLM.2007.009
4. Johnston SS, Morton JM, Kalsekar I, Ammann EM, Hsiao CW, Reps J. Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery. *Value Heal*. 2019;22(5):580-586. doi:10.1016/j.jval.2019.01.011
5. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digit Med*. 2018;1(1):1-10. doi:10.1038/s41746-018-0029-1
6. Samad MD, Ulloa A, Wehner GJ, et al. Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning. *JACC Cardiovasc Imaging*. 2019;12(4):681-689. doi:10.1016/j.jcmg.2018.04.026
7. Pilote L, Raparelli V. Participation of Women in Clinical Trials: Not Yet Time to Rest on Our Laurels. *J Am Coll Cardiol*. 2018;71(18):1970-1972. doi:10.1016/j.jacc.2018.02.069
8. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA Intern Med*. 2014;174(11):1868-1870. doi:10.1001/jamainternmed.2014.4758
9. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals. *JAMA*. 2007;297(11):1233-1240. doi:10.1001/jama.298.1.39-b
10. Garcia M, Mulvagh SL, Merz CNB, Buring JE, Manson JAE. Cardiovascular disease in women: Clinical perspectives. *Circ Res*. 2016;118(8):1273-1293. doi:10.1161/CIRCRESAHA.116.307547
11. Jagannathan R, Patel SA, Ali MK, Narayan KMV. Global Updates on Cardiovascular Disease Mortality Trends and Attribution of Traditional Risk Factors. *Curr Diab Rep*. 2019;19(7). doi:10.1007/s11892-019-1161-2
12. Seidell JC, Halberstadt J. The global burden of obesity and the challenges of prevention. *Ann Nutr Metab*. 2015;66(suppl 2):7-12. doi:10.1159/000375143
13. Gentile BA. Contraindications to Stress Testing. In*: Pocket Guide to Stress Testing*. John Wiley & Sons, Ltd; 2019:45-52. doi:https://doi.org/10.1002/9781119481737.ch4
14. Menger V, Scheepers F, van Wijk LM, Spruit M. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telemat Informatics*. 2018;35(4):727-736. doi:10.1016/j.tele.2017.08.002
15. Siegersma KR, Groepenhoff F, Onland-Moret NC, et al. New York Heart Association class is strongly associated with mortality beyond heart failure in symptomatic women. *Eur Hear J - Qual Care Clin*

Outcomes. 2021;7(2):214-215. doi:10.1093/ehjqcco/qcaa091
16. Groepenhoff F, Eikendal ALM, Charlotte Onland-Moret N, et al. Coronary artery disease prediction in women and men using chest pain characteristics and risk factors: An observational study in outpatient clinics. *BMJ Open*. 2020;10(4). doi:10.1136/bmjopen-2019-035928
17. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *J Biomed Semantics*. 2018;9(1):1-13. doi:10.1186/s13326-018-0179-8
18. Nielen MMJ, Poos MJJC, Baan CA, Gommer AM, Rodriguez M. Prevalentie diabetes in huisartsenpraktijk naar leeftijd en geslacht. Volksgezondheidenzorg.info. Published 2020. https://www.volksgezondheidenzorg.info/onderwerp/diabetes-mellitus/cijfers-context/huidige-situatie
19. Smeets HM, Kortekaas MF, Rutten FH, et al. Routine primary care data for scientific research, quality of care programs and educational purposes: The Julius General Practitioners' Network (JGPN). *BMC Health Serv Res*. 2018;18(1):1-9. doi:10.1186/s12913-018-3528-5

SUPPLEMENTARY MATERIALS

Supplementary table 1 Medication names per medication group

Group	Medications included
Aspirin	Acetylsalicic acid, Carbasalate calcium
Angiotensin-converting-enzyme inhibitor (ACEI)	Benazepril, Perindopril, Captopril, Cilazapril, Delapril, Enalapril, Fosinopril, Lisinopril, Quinapril, Ramipril, Trandolapril, Zofenopril
Angiotensin receptor blocker (ARB)	Candesartan, Eprosartan, Irbesartan, Losartan, Olmesartan, Telmisartan, Valsartan
Thiazides	Hydrochlorothiazide, Chlorthalidone, Indapamide
Potassium-sparing diuretics	Eplerenone, Spironolactone, Triamterene
Loop diuretics	Bumetanide, Furosemide
Beta-blocker	Acebutolol, Atenolol, Bisoprolol, Carvedilol, Celiprolol, Labetalol, Metoprolol, Nebivolol, Pindolol, Propranolol, Sotalol
Calcium-channel blocker	Amlodipine, Barnidipine, Felodipine, Isradipine, , Lacidipine, Lercanidipine, Nicardipine, Nifedipine, Nimodipine, Nitrendipine, Diltiazem, Verapamil
Alpha-blocker	Alfuzosine, Doxazosine, Silodosine, Tamsulosine, Terazosine, Urapidil
Nitrates	Isosorbide dinitrate, , Isosorbide mononitrate, Nicorandil, Nitroglycerine
Digoxin	Digoxin
Statins	Atorvastatine, Fluvastatine, Pitavastatine, Pravastatine, Rosuvastatine, Simvastatine
Metformin	Metformin
Insulin	Insuline
Ezetimibe	Ezetimibe
Sulphonylureas	Glibenclamide, Glimepiride, Tolbutamide, Gliclazide
Fibrates	Bezafibrate, Ciprofibrate, Gemfibrozil, Fenofibrate
P2Y12-Inhibitor	Prasugrel, Clopidogrel, Ticagrelor
Dipyridamole	Dipyridamole
Ivabradine	Ivabradine
Non Vitamin-K oral anticoagulant (NOAC)	Apixaban, Edoxaban, Rivaroxaban, Dabigatran
Anti-arrhythmics	Amiodarone, Disopyramide, Flecainide, Kinidine, Lidocaine, Propafenone
Vitamin-K Antagonist	Acenocoumarol, Phenprocoumon
Other	Any medication that is not in any of the groups described above

Supplementary table 2 Diagnoses per diagnosis group

Group	Diagnoses included
Cardiovascular disease	
Heart failure	Left ventricular hypertrophy, left ventricular dysfunction, concentric hypertrophic left ventricle, concentric left ventricle, decompensatio cordis, heart failure, diastolic dysfunction, coronary microvascular disease, poor ventricular function
Coronary heart disease	Myocardial infarction, angina pectoris, anginal symptoms, chest pain, acute coronary syndrome, silent ischaemia, coronary disease, heart revalidation, coronary insufficiency, 1/2/3 artery disease
Cerebrovascular disease	Cerebrovascular accident, transient ischaemic attack, subarachnoid haemorrhage, eye infarct, brain infarct, brain bleeding, stroke, subarachnoidal bleeding, cerebral infarct, cerebrovascular infarct, retina infarct, lacunar infarct
Congenital heart disease	Tetralogy of Fallot, ventricular septum defect, atrial septum defect, septum defect, coarctatio aortae, foramen ovale, Ductus Botalli
Cardiovascular intervention	Percutaneous coronary intervention, stent, coronary artery bypass graft, bypass, revascularisation, grafting, dotter, percutaneous transluminal coronary angioplasty, valve replacement, valvuloplasty, transcatheter aortic valve implantation, aortic valve replacement, mitral valve replacement, commissurotomy, myocardial perfusion scan, heart catheterisation, implementation of pacemaker or implantable cardioverter-defibrillator
Conditions that are risk factors for cardiovascular disease	
Other cardiovascular disease	Cardiomyopathy, atherosclerosis, abdominal aortic aneurysm, peripheral vascular disease, arteriosclerosis, claudicatio intermittens, deep vein thrombosis, venous thrombosis, venous insufficiency, phlebitis
Arrhythmia	Atrial fibrillation, ventricular fibrillation, atrium flutter, ventricular flutter, paroxysmal atrial fibrillation, conduction delay, supraventricular tachycardia, sick sinus syndrome, sinus exit block, Wolff-Parkinson-White, atrioventricular nodal re-entry tachycardia, extrasystoles, ventricular extrasystoles, arrhythmia, bradycardia, tachycardia, bigemini, AV block, right bundle branch block, left bundle branch block, left anterior hemiblock, premature ventricular contractions, premature atrial contractions, atrial extrasystoles, hemiblock, rhythm disorder
Valvular disease	Valve stenosis, valve sclerosis, valve insufficiency, regurgitation, mitral insufficiency, tricuspid insufficiency, valve defect, mitral regurgitation, mitral stenosis, valve disease
Risk factor intervention	Ablation, radiofrequency catheter ablation, cardioversion, electrocardioversion, percutaneous transluminal angioplasty, endarterectomy, aortic bifurcation prosthesis, abdominal aortic stent

Supplementary table 3 Baseline characteristics stratified by follow-up status

Variable	Whole cohort	Follow-up	No follow-up	Missing data (%)
n	109,151	47,755	61,396	
Women (n, %)	56,628 (51.9)	24,271 (50.8)	32,357 (52.7)	
Age (years)	56 (15)	60 (14)	54 (16)	
Body mass index (kg/m^2)	27.4 (20.0)	27.8 (24.4)	27.0 (15.9)	2.9
Systolic blood pressure (mmHg)	141 (22)	145 (22)	138 (20)	2.9
Current smoker (n, %)	40,139 (36.8)	19,645 (41.1)	20,494 (33.4)	8.9
Ever smoker (n, %)	71,659 (65.7)	34,250 (71.7)	37,409 (60.9)	8.8
Cardiovascular disease (CVD) (n, %)				
History of CVD	16,311 (14.9)	9845 (20.6)	6466 (10.5)	
Family history of CVD	71,148 (65.2)	31,125 (65.2)	40,023 (65.2)	17.8
CVD risk factor conditions	23,957 (21.9)	14,465 (30.3)	9492 (15.5)	
Comorbidities (n, %)				
Hypertension	32,460 (29.7)	17,238 (36.1)	15,222 (24.8)	2.5
Dyslipidaemie	16,978 (15.6)	8765 (18.4)	8213 (13.4)	2.5
Diabetes mellitus	8709 (8.0)	4329 (9.1)	4380 (7.1)	2.6

Supplementary table 4 Baseline characteristics stratified by external referral status

	Whole cohort	External procedure	No external procedure	Missing data (%)
n	109,151	18,050	91,101	
Women (n, %)	56,628 (51.9)	8322 (46.1)	48,306 (53.0)	
Age (years)	56 (15)	60 (12)	56 (16)	
Body mass index (kg/m^2)	27.4 (20.0)	27.7 (15.3)	27.3 (20.8)	2.9
Systolic blood pressure (mmHg)	141 (22)	144 (21)	141 (22)	2.9
Current smoker (n, %)	40,139 (36.8)	6169 (34.2)	33,970 (37.3)	8.9
Ever smoker (n, %)	71,659 (65.7)	11,981 (66.4)	59,678 (65.5)	8.8
Cardiovascular disease (CVD) (n, %)				
History of CVD	16,311 (14.9)	3839 (21.3)	12,472 (13.7)	
Family history of CVD	71,148 (65.2)	12,492 (69.2)	58,656 (64.4)	17.8
CVD risk factor conditions	23,957 (21.9)	4531 (25.1)	19,426 (21.3)	
Comorbidities (n, %)				
Hypertension	32,460 (29.7)	6389 (35.4)	26,071 (28.6)	2.5
Dyslipidaemie	16,978 (15.6)	3583 (19.9)	13,395 (14.7)	2.5
Diabetes mellitus	8709 (8.0)	1864 (10.3)	6845 (7.5)	2.6

Chapter 3

Outcomes in patients with a first episode of chest pain undergoing early coronary CT imaging

Klaske R. Siegersma, N. Charlotte Onland-Moret, Yolande Appelman, Pim van der Harst, Igor I. Tulevski, G. Aernout Somsen, Jagat Narula, Hester M. den Ruijter, Leonard Hofstra

Heart 2022; 108: 1361-1368

ABSTRACT

Objectives To investigate the impact of a CT-first strategy on all-cause and cardiovascular mortality in patients presenting with chest pain in outpatient cardiology clinics.

Methods Patients with a first presentation of suspected angina pectoris were identified and their data linked to the registrations of Statistics Netherlands for information on mortality. The linked database consisted of 33,068 patients. CT-first patients were defined as patients with a CT calcium score and coronary CT angiography, within 6 weeks after their initial visit. Propensity score matching (1:5) was used to match patients with and without a CT-first strategy. After matching, 12,545 patients were included of which 2,308 CT-first patients and 10,237 patients that underwent usual care.

Results Mean age was 57 years, 56.3% were women and median follow-up was 4.9 years. All-cause mortality was significantly lower in CT-first patients (n = 43, 1.9%) compared with patients without CT (n = 363, 3.5%) (hazard ratio: 0.51 [95% CI, 0.37-0.70]). Furthermore, CT-first patients were more likely to receive cardiovascular preventative and anti-anginal medication (aspirin: 44.9% vs 27.1%, statins: 48.7% vs 30.3%, beta-blockers: 37.8% vs 25.5%, in CT-first and without CT-first patients, respectively) and to undergo downstream diagnostics and interventions (coronary interventions: 8.5% vs 5.7%, coronary angiography: 16.2% vs 10.6% in CT-first and without CT-first patients, respectively).

Conclusions In a real-world regular care database, a CT-first strategy in patients suspected of angina pectoris was associated with a lowering of all-cause mortality.

INTRODUCTION

Longitudinal studies of employing cardiac computed tomography (CT) imaging including both coronary calcium scoring (CACS) and cardiac CT angiography (CCTA) in patients presenting with chest pain have demonstrated incremental prognostic value compared to traditional risk profiling algorithms.[1] Higher CACS is associated with mortality[2,3] and has added value to the Framingham Risk Score (FRS) for predicting cardiovascular events.[4] The introduction of CCTA showed even greater promise as a tool to define risk of myocardial infarction[5] and coronary revascularization[6] compared to CACS. Prospectively randomized trials showed that CCTA was superior to functional cardiac testing for cardiovascular endpoints.[7,8]

Until the results of the SCOT-HEART study were published[5], CCTA was only adopted by the guidelines of the National Institute for Health and Care Excellence in the UK[9]. The SCOT-HEART study randomized patients presenting with chest pain in the cardiology outpatient clinic to either CCTA-initiated, the so-called CT-first strategy, or routine clinical care. Patients randomized to the CCTA arm showed only about half of the fatal and non-fatal myocardial infarctions after a 5-year follow-up[5], although these results were not in line with the previously published PROMISE-trial[8] and multiple registry studies[10,11]. Nevertheless, these results prompted the European Society of Cardiology to accord CCTA a class I recommendation for the diagnosis of coronary artery disease (CAD) in symptomatic patients in the renewed guidelines of 2019.[12]

However, the results of SCOT-HEART might not be generalizable to a routine care population who are expected to have a higher likelihood of mortality, since randomized trials may induce an overestimation of benefit due to the so-called healthy volunteer inclusion bias in trials.[13,14] Furthermore, multiple patient groups are often underrepresented in clinical studies, such as women[15], the elderly, non-white ethnicities, and patients with comorbidities.[16,17] Therefore, the impact of a CT-first strategy for patients with chest pain in regular care is still unknown.[10,11]

The use of data from real world databases to assess the utility of cardiac CT may aid in defining associations between patient variables and outcomes in the general population. Moreover, real world data may help to vindicate results of clinical trials, especially with the inclusion of substantial numbers of all relevant patient groups. Thus, we used a regular care database[18] to investigate all-cause and cardiovascular mortality in patients suspected of angina pectoris, who underwent CT calcium score and CCTA following their first visit, and compared this with a propensity score (PS) matched control group from the same database, who were subjected to regular care without cardiac CT.

METHODS

Study population

The CCN database consists of 109,151 patients who visited one of the CCN's outpatient cardiology clinics between 2007 and 2018. These centers are known for their homogenized and structured approach to investigate patients with cardiovascular complaints. A detailed description of the CCN database has been published.[18] Patients with chest pain suspected to be angina pectoris, who presented for the first time at one of the diagnostic centers were included. The study population was then split in two groups; patients with a cardiac CT within six weeks after their first visit, the CT-first strategy, and patients without this diagnostic procedure. Cardiac CT included a scanning protocol for calcium scoring and CCTA. The decision for referral for cardiac CT was made by the treating cardiologist. The type of CT-scanner was based on availability in the referenced centers. Follow-up of the population was obtained through linking with the population database of Statistics Netherlands (CBS). Figure 1 outlines the selection of the study population.

The Cardiology Centers of the Netherlands data were made available under implied consent and transferred to the University Medical Center Utrecht under the Dutch Personal Data Protection Act. This study used data collected during the regular care process and did not subject participants to additional procedures or impose behavioural patterns on them. The Medical Research Ethics Committee of the University Medical Center Utrecht declared that research within the CCN database does not meet the Dutch Medical Research Involving Human Subjects Act (proposal number 17/359).

Study variables

Patient characteristics, prior co-morbidities, risk factors, family history and a general medical history were obtained from the electronic health records (EHR) of the patient. Medication use was extracted from pharmaceutical prescription data. Residential region of the patient was obtained through 4-digit postal code. Chest pain was characterized as typical, atypical or non-anginal, according to Diamond[12,19] and obtained through retrieval of text variables in the EHR. We also included the results from the stress electrocardiogram (ECG) recording to ensure comparability between patients with and without a CT-first strategy.

Short- and long-term outcome

All-cause mortality and cardiovascular mortality were the primary and secondary outcome in the analyses, respectively. Other secondary outcomes included registration of diagnostic and therapeutic procedures in the EHR. Diagnostics included stress ECG recording, functional imaging (cardiac magnetic resonance imaging [MRI], positron emission tomography [PET], and single-photon emission computed tomography [SPECT]),

invasive coronary angiography (CAG) and cardiac CT more than six weeks after the first visit to CCN. Therapeutic procedures were percutaneous coronary intervention (PCI) and coronary artery bypass grafting (CABG). Moreover, prescribed medication changes during and after the chest pain consult were evaluated and compared between the groups. This analysis focused primarily on the value of a cardiac CT in the diagnostic trajectory of the patient. We did not evaluate the association between results of cardiac CT and primary and secondary outcomes.

Missing data

Missing data in variables required for analysis were handled by multiple imputation for chained equations (MICE; 10 iterations, 10 imputed datasets) with the R-package mice (version 3.8.0).[20] However, for the presence of a family history, missing data were not at random. Therefore, missing values for this variable were filled with a negative family history. Included variables for multiple imputation models are indicated in Table 1.

Propensity score matching

PS matching[21] was used to ensure a comparable sample between patients with and without a CT-first strategy. PSs were calculated for each patient separately in each imputed dataset with a logistic regression model including 19 variables, as indicated in Table 1. This selection of variables was based upon clinically relevant variables and baseline differences between both groups. Thereafter, the average PS of the imputation datasets was calculated for each patient. We matched 5 patients who did not receive a CT-first strategy to 1 patient with a CT-first strategy based upon the calculated PSs per patient using the nearest neighbour method with a calliper width of 0.05 and no replacement (R-package: matchit[22], version 3.0.2). Comparability of the groups after matching was assessed by inspection of the balance of baseline variables. The selected patient population is further referred to as the matched sample.

Statistical analysis

Descriptive statistics are presented as mean with standard deviation (SD) or median with interquartile range (IQR), where appropriate, for continuous variables and counts and percentages for categorical variables. All-cause and cardiovascular mortality were analysed with Kaplan-Meier curves and Cox regression models on the matched patient selection. Other secondary outcomes were compared with chi-square testing.

Subgroup analyses were performed for type of chest pain (anginal/non-anginal), sex (men/women), age (<65 year/≥ 65 years), SCORE (<5%/≥5%)[23] and pre-test probability of CAD (<5%/>5%)[24]. The p-value for interaction was determined for each analysis. Outcome for all subgroup analyses was all-cause mortality.

Sensitivity analyses were done to evaluate the effect of patient's residency and inclusion

Table 1 Baseline table representing the distribution of baseline variables before and after matching on propensity score.

	Original Sample			Matched Sample			Missing
	Overall	Patients without a CT-first strategy	CT-first strategy	Overall	Patients without a CT-first strategy	CT-first strategy	%
n	33,068	30,756	2312	12,545	10,237	2308	
Age (mean (SD)) *†	56 (13.41)	55 (13.63)	57 (9.85)	57 (12.62)	57 (13.16)	57 (9.85)	0.0
Female (n, %) *†	17,622 (53.3)	16,329 (53.1)	1293 (55.9)	7068 (56.3)	5779 (56.5)	1289 (55.8)	0.0
Hypertension (n, %)*†	9668 (29.4)	8966 (29.3)	702 (30.4)	3920 (31.2)	3220 (31.5)	700 (30.3)	0.5
Dyslipidaemia (n, %) *†	5475 (16.6)	5084 (16.6)	391 (16.9)	2127 (17.0)	1737 (17.0)	390 (16.9)	0.5
Diabetes (n, %) *†	2734 (8.3)	2559 (8.4)	175 (7.6)	978 (7.8)	803 (7.8)	175 (7.6)	0.6
Height (mean (SD)) *†	173 (9.99)	173(9.99)	173 (9.96)	173 (9.92)	173 (9.91)	173 (9.96)	1.1
Weight (mean (SD)) *†	80 (16.13)	80 (16.15)	80 (15.86)	80 (16.30)	80 (16.40)	80 (15.86)	1.1
BMI (mean (SD)) *†	27 (4.71)	27 (4.72)	27 (4.60)	27 (4.75)	27 (4.78)	27 (4.61)	1.2
Chest pain category (n, %) *†							56.4
Non-anginal	8198 (56.8)	7723 (58.5)	475 (39.0)	1965 (36.7)	1490 (36.1)	475 (39.1)	
Atypical	2092 (14.5)	1906 (14.4)	186 (15.3)	861 (16.1)	675 (16.3)	186 (15.3)	
Typical	4131 (28.6)	3574 (27.1)	557 (45.7)	2521 (47.1)	1967 (47.6)	554 (45.6)	
Smoking status (n, %) *†							6.3
Current	13,050 (42.1)	12,285 (42.6)	765 (35.5)	4067 (35.0)	3303 (34.9)	764 (35.5)	
Former	9907 (32.0)	9156 (31.7)	751 (34.9)	4091 (35.2)	3342 (35.3)	749 (34.8)	
Never	8041 (25.9)	7404 (25.7)	637 (29.6)	3456 (29.8)	2819 (29.8)	637 (29.6)	
Family history of atherosclerosis (n, %) *†	11,644 (35.2)	10,593 (34.4)	1051 (45.5)	5662 (45.1)	4615 (45.1)	1047 (45.4)	0.0
Consult year (n, %) *†							0.0
2007-2010	4344 (13.1)	4319 (14.0)	25 (1.1)	133 (1.1)	108 (1.1)	25 (1.1)	
2011-2014	14,362 (43.4)	13,234 (43.0)	1128 (48.8)	6025 (48.0)	4900 (47.9)	1125 (48.7)	
2015-2018	14,362 (43.4)	13,203 (42.9)	1159 (50.1)	6387 (50.9)	5229 (51.1)	1158 (50.2)	
Diagnosis of CHD at baseline (n, %) *†	2731 (8.3)	2612 (8.5)	119 (5.1)	659 (5.3)	540 (5.3)	119 (5.2)	0.0

Diagnosis of cerebrovascular disease at baseline (%) *†	975 (2.9)	918 (3.0)	57 (2.5)	303 (2.4)	246 (2.4)	57 (2.5)	0.0
Medication prescribed (n, %) *†							0.0
Aspirin	1209 (3.7)	1130 (3.7)	79 (3.4)	442 (3.5)	363 (3.5)	79 (3.4)	
Betablocker	1266 (3.8)	1171 (3.8)	95 (4.1)	537 (4.3)	442 (4.3)	95 (4.1)	
Calcium channel blocker	608 (1.8)	572 (1.9)	36 (1.6)	184 (1.5)	148 (1.4)	36 (1.6)	
Nitrate	522 (1.6)	498 (1.6)	24 (1.0)	124 (1.0)	100 (1.0)	24 (1.0)	
Statin	1335 (4.0)	1259 (4.1)	76 (3.3)	415 (3.3)	339 (3.3)	76 (3.3)	
Conclusion of exercise test (n, %) *†							17.7
Abnormal	2295 (8.4)	2295 (8.4)	346 (18.7)	1570 (15.6)	1227 (15.0)	343 (18.6)	
Inconclusive	5622 (20.7)	5622 (20.7)	509 (27.5)	2960 (29.5)	2452 (29.9)	508 (27.5)	
Normal	19,308 (70.9)	19,308 (70.9)	997 (53.8)	5518 (54.9)	4521 (55.1)	997 (54.0)	
Domestic region in the NL (n, %) *†							0.0
Middle	13,991 (42.3)	12,693 (41.3)	1298 (56.1)	7103 (56.6)	5809 (56.7)	1294 (56.1)	
North	17,339 (52.4)	16,344 (53.1)	995 (43.0)	5332 (42.5)	4337 (42.4)	995 (43.1)	
South	1738 (5.3)	1719 (5.6)	19 (0.8)	110 (0.9)	91 (0.9)	19 (0.8)	
Total cholesterol in mmol/L (mean (SD)) *†	5.15 (1.13)	5.13 (1.13)	5.33 (1.17)	5.32 (1.16)	5.31 (1.16)	5.33 (1.16)	24.4
10-year HeartSCORE in % (median [IQR])	1.75 [0.47-4.97]	1.73 [0.44-5.03]	1.92 [0.75-4.46]	2.04 [0.60-5.45]	2.08 [0.57-5.67]	1.91 [0.75-4.47]	30.9
Pretest probability (median [IQR])	10 [3-22]	10 [3-22]	13 [6-22]	13 [6-26]	13 [6-32]	13 [6-22]	61.8
Pretest probability of CAD in % (n, %)							61.8
High >15%	4412 (34.9)	3939 (34.2)	473 (41.7)	2128 (45.0)	1657 (46.0)	471 (41.6)	
Intermediate 5%-15%	4028 (31.9)	3600 (31.3)	428 (37.7)	1512 (31.9)	1085 (30.1)	427 (37.8)	
Low <5%	4199 (33.2)	3966 (34.5)	233 (20.5)	1093 (23.1)	860 (23.9)	233 (20.6)	
All-cause mortality (n, %)	1331 (4.0)	1288 (4.2)	43 (1.9)	406 (3.2)	363 (3.5)	43 (1.9)	0.0
Cardiovascular mortality (n, %)	329 (1.0)	313 (1.0)	16 (0.7)	111 (0.9)	95 (0.9)	16 (0.7)	0.0
Follow-up in years (median [IQR])	5.5 [3.4-7.8]	5.6 [3.4-7.9]	5.1 [3.3-6.9]	4.9 [3.2-6.7]	4.9 [3.2-6.7]	5.1 [3.3-6.8]	0.0
Nelson-Aalen estimator* (median [IQR])	0.035 [0.019-0.058]	0.035 [0.019-0059]	0.031 [0.019-0.047]	0.030 [0.018-0.046]	0.030 [0.018-0.046]	0.031 [0.019-0.047]	0.0

CVD: cardiovascular disease, BMI: Body Mass Index, IQR: interquartile range, SD: standard deviation. Pre-test probability is derived from the 2019, ESC Guidelines[12,24].The summary statistics are shown as the data before imputation. Missing are the percentage of missing observations in the total population. *Variables used for multiple imputation, †Variables used for calculation of PS.

year. In these analyses a Cox regression for all-cause mortality was performed. Another sensitivity analysis excluded all patients referred to one specific diagnostic center, only performing CCTA in case of medical need or if calcium score was above 0. As no information of the center that was visited is available in the database, residency of the patient is used as a substitute. A final sensitivity analysis removed all patients without a CT, but with CAG to evaluate the effect on the results. These sensitivity analyses evaluated Cox regressions for both cardiovascular and all-cause mortality. All data analyses were done with R (version 3.6.2) and RStudio (version 1.1.463).

Patient and Public involvement

Patients were not involved in any stage of this research process.

RESULTS

Study population

A total of 34,311 patients with chest pain met the inclusion criteria and were selected from the CCN database (figure 1). After linking with the database of Statistics Nether-

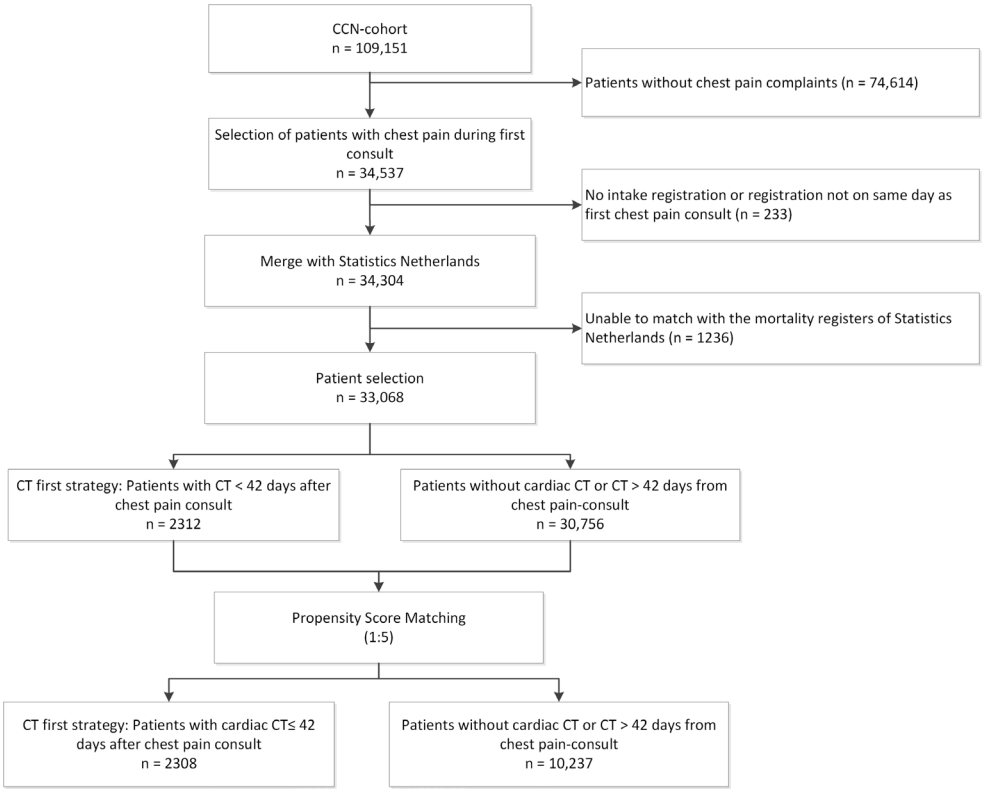


Figure 1 Flowchart of patient selection

lands, mortality data of 33,068 (96.4%) patients were available. This selection included 2,312 and 30,756 patients, respectively, with and without a CT-first strategy. Mean age of the included patients (n=33,068) was 56 years and 53.3% were women. Before PS matching, patients in the CT-first group (table 1) were older (57 vs 55 years), less likely to be current smokers (35.5% vs 42.6%) and had a higher incidence of typical chest pain (45.7% vs 27.1 %). Yet, the presence of comorbidities was comparable. Median follow-up was 5.5 [IQR 3.4-7.8] years, and 1331 (4.0%) patients died during follow-up. In 329 (1.0%) individuals a cardiovascular cause of death was assigned. Thus, PS were distributed differently between patients with and without a CT-first strategy (Supplementary figure 1).

After PS matching, 2,308 patients were in the CT-first group and 10,237 patients were matched to this group. Patients in the CT-first group had a median coronary artery calcification score of 2.20 [IQR 0-88]. There were 871, 329 and 225 filed stenoses in the LAD, RCA and CX, respectively, of which 489 (34.3%) were reported to be significant (>50% stenosis). In the matched sample, baseline characteristics were more equally distributed as compared to the unmatched sample (Table 1), e.g. mean age (57 years in both groups) and type of chest pain (non-anginal 39.1% vs 36.1%, atypical 15.3% vs 16.3%, and typical

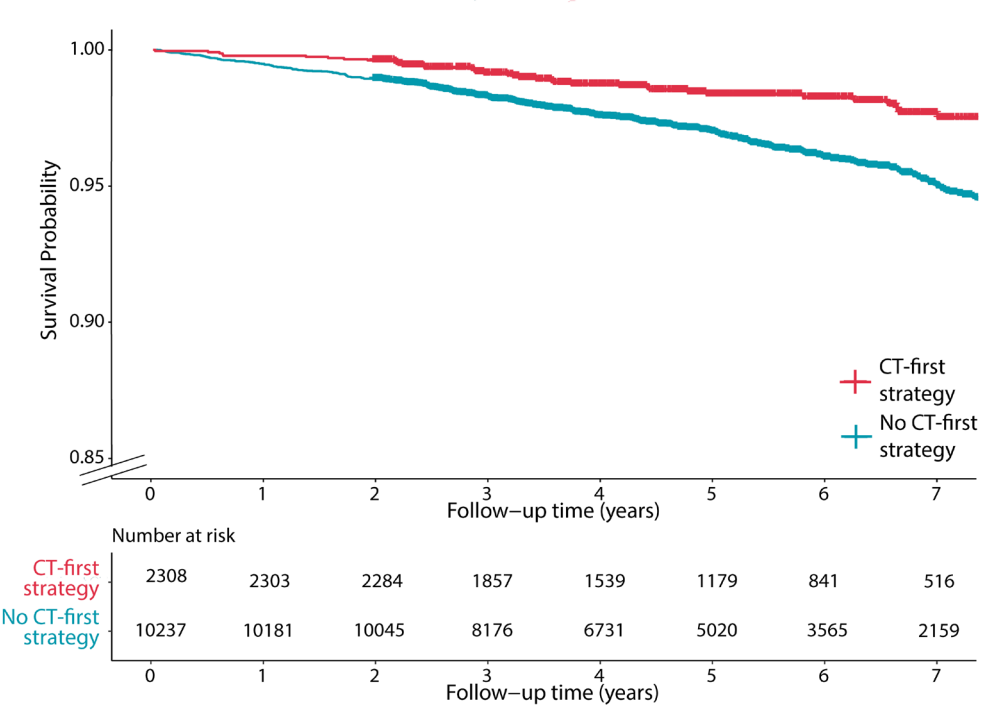


Figure 2 Kaplan-Meier curves for all-cause mortality of the CT-first and the without a CT-first study population.

Table 2 Distribution of diagnostics and therapeutics during follow-up in patients with a CT-first strategy and the population without a CT-first strategy.

	CT-first (n=2308)	Patients without a CT-first strategy (n=10,237)	p-value
Anatomical imaging (n, %)	392 (17.0)	1435 (14.0)	<.001
Perfusion imaging (n, %)	70 (3.0)	290 (2.8)	.652
Coronary interventions (n, %)	197 (8.5)	581 (5.7)	<.001
Stress ECG at CCN (n, %)	403 (17.5)	1600 (15.6)	.032
Coronary Angiography (n, %)	373 (16.2)	1086 (10.6)	<.001
Cardiac CT (n, %)	23 (1.0)	429 (4.2)	<.001
Time difference between chest pain consult and intervention			
Anatomical imaging	57 [37.75-87.25]	49 [20-119.50]	<.001
Perfusion imaging	74 [49.25-162.25]	36.50 [17-91]	<.001
Coronary intervention	65 [42-98]	51 [21-102]	<.001
Stress ECG	301 [91-759]	112 [25-597.75]	<.001

Anatomical imaging comprises cardiac CT and coronary angiography. Perfusion imaging includes cardiac PET, SPECT and MRI. Coronary interventions are coronary artery bypass grafts (CABG) and percutaneous coronary interventions (PCI).

45.6% vs 47.6% for, respectively, patients with and without a CT-first strategy). Consequently, the distribution of the PS of the included patients showed improved overlap in the matched sample (Supplementary Figure 1). The matched sample (n=12,545) was used for the analysis of primary and secondary outcomes.

Outcome analysis

All-cause and cardiovascular mortality after PS matching

In the matched sample 406 (3.2%) patients died. In 111 (0.9%) patients there was a cardiovascular cause of death. Median follow-up was 4.9 years. In patients with a CT-first strategy, all-cause mortality was 1.9% (n=43) compared to 3.5% (n=363) in patients without a CT-first strategy (hazard ratio, HR 0.51, 95% CI 0.37-0.70). The Kaplan-Meier curves for patients with and without a CT-first strategy are shown in figure 2 and show a consistent divergent pattern over up to 7 years of follow-up. Cardiovascular mortality for patients in the CT-first group was 16 (0.7%) and 95 (0.9%) for the patients without a CT-first strategy, respectively. The corresponding HR for cardiovascular mortality was 0.73 (95% CI 0.43-1.24). The corresponding Kaplan-Meier curves are shown in Supplementary Figure 2.

Downstream diagnostics and interventions

A total of 3,432 (27.4%) patients of the matched cohort had diagnostic or therapeutic follow-up. The percentage of CAGs (16.2% vs 10.6%, p<.001) and coronary interventions

Table 3 Medication use in selected patients for the CT-first strategy and patients without a CT-first strategy.

	CT-first (n = 2308)	Patients without a CT-first strategy (n = 10,237)
Aspirin (n, %)		
Continued	57 (2.5)	323 (3.2)
Discontinued	22 (1.0)	40 (0.4)
Initiated	1036 (44.9)	2770 (27.1)
Initiated and discontinued	339 (14.7)	369 (3.6)
Beta-blocker (n, %)		
Continued	75 (3.2)	419 (4.1)
Discontinued	20 (0.9)	23 (0.2)
Initiated	873 (37.8)	2611 (25.5)
Initiated and discontinued	549 (23.8)	525 (5.1)
Calcium-channel blocker (n, %)		
Continued	36 (1.6)	135 (1.3)
Discontinued	<10 (<0.4)	13 (0.1)
Initiated	287 (12.4)	1311 (12.8)
Initiated and discontinued	81 (3.5)	213 (2.1)
Nitrates (n, %)		
Continued	18 (0.8)	92 (0.9)
Discontinued	<10 (<0.4)	<10 (<0.1)
Initiated	502 (21.8)	1791 (17.5)
Initiated and discontinued	165 (7.1)	260 (2.5)
Statins (n, %)		
Continued	71 (3.1)	332 (3.2)
Discontinued	<10 (<0.4)	<10 (<0.1)
Initiated	1125 (48.7)	3102 (30.3)
Initiated and discontinued	84 (3.6)	141 (1.4)

Continued medication is defined as medication that was started before the chest pain consult and continued for at least 120 following days. Discontinued medication is medication started before the consult and stopped within 120 days. Initiated medication was started within the timeframe from chest pain consult until 120 days after the consult. Initiated and discontinued medication was medication that was started and discontinued within 120 days following the chest pain consult.

(8.5% vs 5.7%, p<.001) were higher in the CT-first strategy group, compared to the group of patients without CT (Table 2). The same was seen for the number of stress ECGs, which was higher for CT-first patients (17.5% vs 15.6%, p=.03), yet the time interval between initial presentation and stress ECG was shorter in patients without a CT-first strategy (median time between chest pain consult and stress ECG: 112 days [IQR: 25-598] vs 301 days [IQR: 91-759]). Follow-up cardiac perfusion imaging did not show any differences (3.0% vs 2.8% in, respectively, patients with and without a CT-first strategy).

Medication use

Medication use for primary or secondary prevention of CAD (including aspirin, statins, beta-blockers, calcium channel blockers and nitrates) at baseline was similar for patients

with and without a CT-first strategy (Table 1 and Table 3). After initial chest pain consult, initiation of aspirin (44.9% vs 27.1%, p<.001), betablockers (37.8% vs 25.5%, p<.001), and statins (48.7% vs 30.3%, p<.001) was higher in the CT-first strategy group, compared to the group without CT.

Subgroup and sensitivity analyses

The association between having a cardiac CT for diagnostic evaluation of chest pain and all-cause mortality was further investigated in clinically relevant subgroups (Figure 3). Men showed greater benefit from a CT-first strategy with respect to all-cause mortality, although p-value for interaction was not significant. This p-value was only significant for type of chest pain (p-value: .013) and risk of CAD (p-value: .0046), implying that patients with a possible anginal origin of chest pain and with an intermediate to high risk of CAD benefit from a CT-first strategy.

Sensitivity analyses showed that HRs were similar for the different regions of the Netherlands and for the inclusion year of patients (Supplementary table 1). Residence and inclusion year were taken into account during the PS calculation. This suggests a similar distribution of these variables in the CT-first and without CT-first group. Sensitivity analysis of removal of patients (n=4,883) from one center that did not structurally perform cardiac CT including calcium scoring and CCTA, showed comparable HR for all-cause mortality (0.44, 95% CI 0.28-0.67 vs 0.51, 95% CI 0.37-0.70) and HR for cardiovascular mortality (0.70, 95% CI 0.37-1.33 vs 0.73 95% CI 0.43-1.23) to the main matched sample. After removal of all patients with CAG within 10 weeks after their initial visit (n=959), the HR

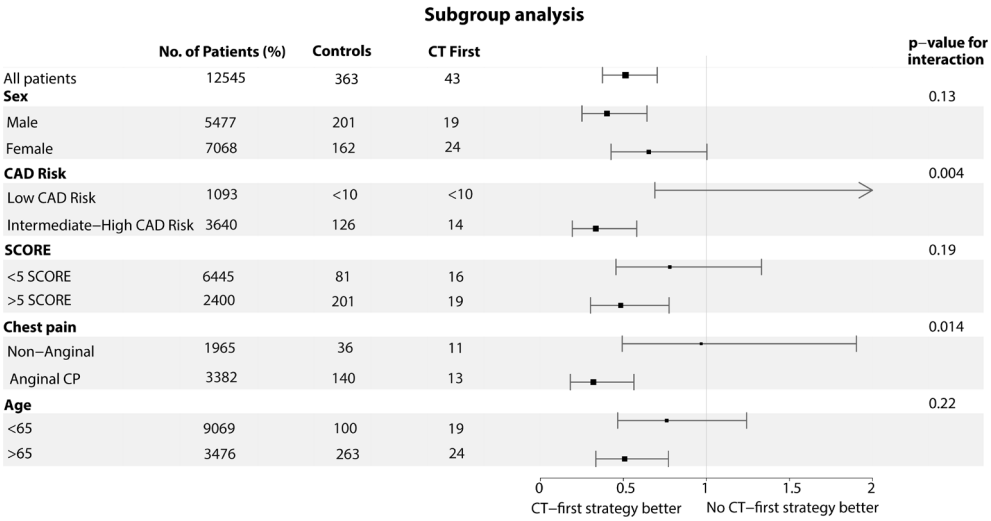


Figure 3 Forest plot of the subgroup analysis. CAD risk is defined as the pre-test probability of coronary artery disease according to the ESC guidelines of 2019: Low < 5%, Intermediate-High >5%. Anginal chest pain includes typical and atypical chest pain. SCORE is the cardiovascular risk score as defined by the Systematic COronary Risk Evaluation.

for all-cause mortality (0.54, 95% CI 0.39-0.74) and cardiovascular mortality (0.78, 95% CI 0.45-1.34) changed, but within confidence intervals.

DISCUSSION

The analysis performed in this study demonstrated that upstream inclusion of cardiac CT in diagnostic work-up of patients with chest pain was associated with a significant lowering of all-cause mortality, irrespective of the outcome of the cardiac CT. We also observed a lower cardiovascular mortality rate in CT-first patients, although this difference was not statistically significant. This can be due to the low number of a cardiovascular cause of death and insufficient registration of the cause of death. Our data also show that patients with a CT-first strategy had more downstream therapeutics and diagnostic testing. This is in accordance with findings from other registries.[10] The presented results underscore and complement the prominent role of cardiac CT in the diagnostic work-up of patients presenting with chest pain in clinical guidelines[12], which has also been internationally addressed.[25]

Based on previous publications, it is assumed that patients that undergo cardiac CT in addition to regular care receive an accurate diagnosis earlier and more often[26], after which targeted anti-anginal and preventative therapy is initiated, resulting in overall superior survival in these patients.[27,28] The prescription of preventative medication, especially statins, aspirin and beta blockers to patients with CT-verified atherosclerotic lesions, was thought to be the major reason for lower mortality rates observed in patients undergoing cardiac CT in the SCOT-HEART trial. However, cardiovascular mortality rates were not significantly lower in the CT-first population. One reason for the lower risk of all-cause mortality could be the identification of relevant extracardiac findings, which occur in approximately 3%. Of these, pulmonary nodules make up the largest share (1.3%).[29] Also the inclusion of patients at lower risk for CAD in the CT-first group, that have not been properly identified due to missing values, might account for the discrepancy in all-cause and cardiovascular mortality after a CT-first strategy. Finally, the number of events and follow-up time might be too low to establish the effect of lifestyle changes as a result of the visualization of coronaries on CT imaging. Yet, as we did not include an in-depth analysis of the results of cardiac CT, we could not relate this to the presence of CAD or atherosclerotic lesions.

In accordance with the SCOT-HEART trial, we observed a higher referral rate for follow-up diagnostics and interventions in CT-first patients.[5,27] We did not perform analyses of these downstream costs. Therefore, no statements can be given about the cost-to-benefit ratio in patients with and without a cardiac CT-first strategy. Yet, in the light of a faster and more often correct diagnosis[27], lower cumulative diagnostic expenses could be expected

for patients with a cardiac CT[26]. Nevertheless, cost-effectiveness of the CT-first strategy needs to be established and should take into account the higher number of revascularizations and medication prescriptions in CT-first patients.[10,30] On the contrary, the relatively low cost and wide availability of CT imaging makes it accessible for more patients compared to other cardiac imaging strategies.

Strengths and limitations

The presented study has multiple strengths. First, the use of EHR data aids to address concerns of health volunteer bias and underrepresentation of patient subgroups that are mostly seen in clinical trials. This is represented in the large study population and equal distribution of sex within the population. Thus, the included population is an actual representation of the population that the guidelines are intended for. Therefore, the results have high external validity. Second, the longer follow-up in this study has incremental value compared to the Danish nationwide registry[10] and the PROMISE trial[8], that presented data with a follow-up of respectively 3.6 years and 25 months.

The study was limited by the lack of information on cardiovascular events, including cardiovascular hospitalizations, which hampered the analysis of these events. Another limitation of regular care data is confounding by indication; patients with a higher pre-test probability of CAD are more likely to be referred for cardiac CT than patients with a lower pre-test probability. To avoid this potential bias, PS matching was used to ensure comparability between patients with and without a CT-first strategy. Regardless, one would have expected that patients with a higher pre-test likelihood would show lower survival. Yet, we observed the opposite, vindicating that baseline differences between both groups were properly accounted for. Nonetheless, it is impossible to eliminate or to take into account the influence of unmeasured confounders or instrumental variables in the presented methodology.

CONCLUSION

To conclude, this study was the first to demonstrate a significantly lower HR for all-cause mortality in patients with chest pain who had a CT-first strategy compared to patients who did not have a CT-first strategy. It is hypothesized that patients with a CT-first strategy obtain a more tailored therapy, including risk-reducing medication. These results support a CT-first strategy for patients with chest pain and strengthen the prominent role for cardiac CT as the primary method for diagnostic work-up of patients with chest pain, as suggested by the current ESC guidelines.[12]

REFERENCES

1. Arad Y, Spadaro LA, Roth M, Newstein D, Guerci AD. Treatment of asymptomatic adults with elevated coronary calcium scores with atorvastatin, vitamin C, and vitamin E: The St. Francis heart study randomized clinical trial. *J Am Coll Cardiol*. 2005;46(1):166-172. doi:10.1016/j.jacc.2005.02.089
2. Budoff MJ, Shaw LJ, Liu ST, et al. Long-Term Prognosis Associated With Coronary Calcification. Observations From a Registry of 25,253 Patients. *J Am Coll Cardiol*. 2007;49(18):1860-1870. doi:10.1016/j.jacc.2006.10.079
3. Detrano R, Guerci AD, Carr JJ, et al. Coronary Calcium as a Predictor of Coronary Events in Four Racial or Ethnic Groups. *N Engl J Med*. 2008;358(13):1336-1345. doi:10.1056/nejmoa072100
4. Yeboah J, McClelland RL, Polonsky TS, et al. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. *JAMA*. 2012;308(8):788-795. doi:10.1001/jama.2012.9624
5. The SCOT-HEART Investigators. Coronary CT Angiography and 5-Year Risk of Myocardial Infarction. *N Engl J Med*. 2018;379(10):924-933. doi:10.1056/nejmoa1805971
6. Villines TC, Hulten EA, Shaw LJ, et al. Prevalence and severity of coronary artery disease and adverse events among symptomatic patients with coronary artery calcification scores of zero undergoing coronary computed tomography angiography: Results from the CONFIRM (Coronary CT Angiography Evalu. *J Am Coll Cardiol*. 2011;58(24):2533-2540. doi:10.1016/j.jacc.2011.10.851
7. Neglia D, Rovai D, Caselli C, et al. Detection of Significant Coronary Artery Disease by Noninvasive Anatomical and Functional Imaging. *Circ Cardiovasc Imaging*. 2015;8(3):1-10. doi:10.1161/CIRCIMAGING.114.002179
8. Douglas PS, Hoffmann U, Patel MR, et al. Outcomes of anatomical versus functional testing for coronary artery disease. *N Engl J Med*. 2015;372(14):1291-1300. doi:10.1056/NEJMoa1415516
9. National Institute for Health and Clinical Excellence. Recent-onset chest pain of suspected cardiac origin: assessment and diagnosis. *Clin Guidel* [CG95]. 2010;(November).
10. Jørgensen ME, Andersson C, Nørgaard BL, et al. Functional Testing or Coronary Computed Tomography Angiography in Patients With Stable Coronary Artery Disease. *J Am Coll Cardiol*. 2017;69(14):1761-1770. doi:10.1016/j.jacc.2017.01.046
11. Roifman I, Wijeysundera HC, Austin PC, Rezai MR, Wright GA, Tu J V. Comparison of anatomic and clinical outcomes in patients undergoing alternative initial noninvasive testing strategies for the diagnosis of stable coronary artery disease. *J Am Heart Assoc*. 2017;6(7):1-13. doi:10.1161/JAHA.116.005462
12. Knuuti J, Wijns W, Achenbach S, et al. 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes. *Eur Heart J*. 2020;41(3):407-477. doi:10.1093/eurheartj/ehz425
13. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am J Epidemiol*. 2007;165(8):874-881. doi:10.1093/aje/kwk075
14. Leening MJG, Heeringa J, Deckers JW, et al. Healthy volunteer effect and cardiovascular risk. *Epidemiology*. 2014;25(3):470-471. doi:10.1097/EDE.0000000000000091
15. Pilote L, Raparelli V. Participation of Women in Clinical Trials: Not Yet Time to Rest on Our Laurels. *J Am Coll Cardiol*. 2018;71(18):1970-1972. doi:10.1016/j.jacc.2018.02.069

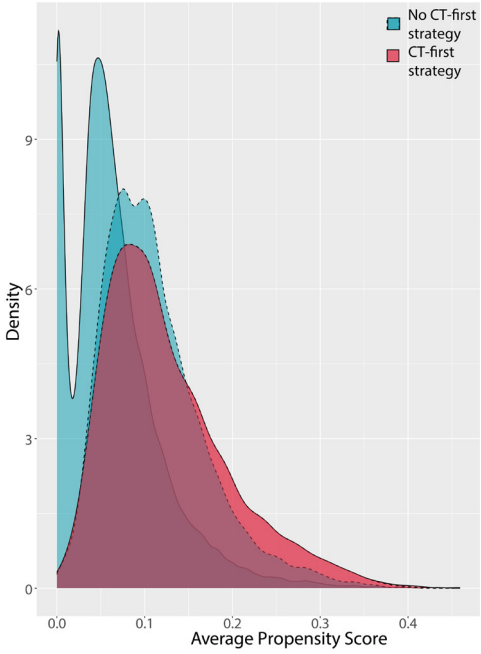
16. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA Intern Med*. 2014;174(11):1868-1870. doi:10.1001/jamainternmed.2014.4758
17. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals. *JAMA*. 2007;297(11):1233-1240. doi:10.1001/jama.298.1.39-b
18. Bots SH, Siegersma KR, Onland-Moret NC, et al. Routine clinical care data from thirteen cardiac outpatient clinics: design of the Cardiology Centers of the Netherlands (CCN) database. *BMC Cardiovasc Disord*. 2021;21(1):1-9. doi:10.1186/s12872-021-02020-7
19. Diamond GA. A clinically relevant classification of chest discomfort. *J Am Coll Cardiol*. 1983;1(2):574-575. doi:10.1016/S0735-1097(83)80093-X
20. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1-67. doi:10.18637/jss.v045.i03
21. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786
22. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42(8):1-28. doi:10.18637/jss.v042.i08
23. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J*. 2003;24(11):987-1003. doi:10.1016/S0195-668X(03)00114-3
24. Juarez-Orozco LE, Saraste A, Capodanno D, et al. Impact of a decreasing pre-test probability on the performance of diagnostic tests for coronary artery disease. *Eur Heart J Cardiovasc Imaging*. 2019;20(11):1198-1207. doi:10.1093/ehjci/jez054
25. Poon M, Lesser JR, Biga C, et al. Current Evidence and Recommendations for Coronary CTA First in Evaluation of Stable Coronary Artery Disease. *J Am Coll Cardiol*. 2020;76(11):1358-1362. doi:10.1016/j.jacc.2020.06.078
26. Lubbers M, Dedic A, Coenen A, et al. Calcium imaging and selective computed tomography angiography in comparison to functional testing for suspected coronary artery disease: The multicentre, randomized CRESCENT trial. *Eur Heart J*. 2016;37(15):1232-1243. doi:10.1093/eurheartj/ehv700
27. Williams MC, Hunter A, Shah ASV, et al. Use of Coronary Computed Tomographic Angiography to Guide Management of Patients with Coronary Disease. *J Am Coll Cardiol*. 2016;67(15):1759-1768. doi:10.1016/j.jacc.2016.02.026
28. Newby D, Williams M, Hunter A, et al. CT coronary angiography in patients with suspected angina due to coronary heart disease (SCOT-HEART): An open-label, parallel-group, multicentre trial. *Lancet*. 2015;385(9985):2383-2391. doi:10.1016/S0140-6736(15)60291-4
29. Karius P, Lembcke A, Sokolowski FC, et al. Extracardiac findings on coronary computed tomography angiography in patients without significant coronary artery disease. *Eur Radiol*. 2019;29(4):1714-1723. doi:10.1007/s00330-018-5688-4
30. Zeb I, Abbas N, Nasir K, Budoff MJ. Coronary computed tomography as a cost-effective test strategy for coronary artery disease assessment - A systematic review. *Atherosclerosis*. 2014;234(2):426-435. doi:10.1016/j.atherosclerosis.2014.02.011

SUPPLEMENTARY MATERIALS

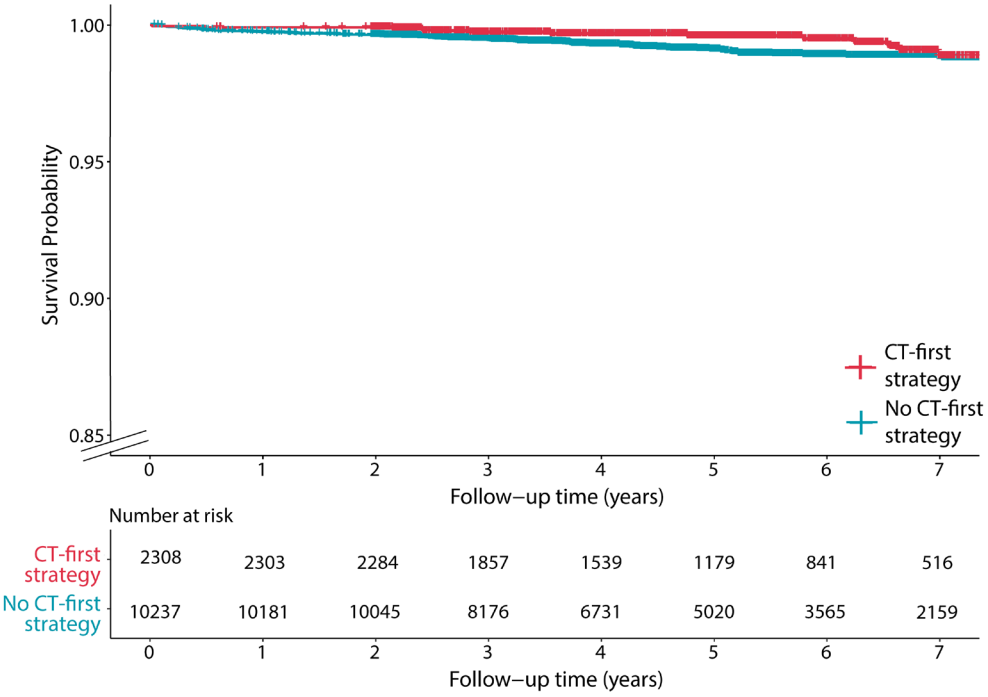
Supplementary table 1 Results of sensitivity analyses. This table shows the results of the cox regression analysis for all-cause mortality for patient's residence of different regions of the Netherlands and for the inclusion year of the patient.

	Number of patients	Without CT-first strategy	CT-first strategy	Events in patients without CT-first strategy	Events in patients with CT-first strategy	Hazard ratio all-cause mortality [95% CI]
All patients	12,545	10,237	2308	363	43	0.51 [0.37-0.70]
Region						
North	5332	4337	995	154	14	0.39 [0.23-0.68]
Middle	7103	5809	1294	206	29	0.60 [0.40-0.88]
South	110	91	19	<10	<10	NA
Year of inclusion						
2007-2010	133	108	25	<10	<10	2.18 [0.39-11.9]
2011-2014	6025	4900	1125	257	30	0.49 [0.34-0.72]
2015-2018	6387	5229	1158	102	11	0.48 [0.25-0.89]

North: provinces of Friesland, Groningen, Drenthe, Overijssel, Flevoland and North-Holland, Middle: provinces of Zuid-Holland, Utrecht and Gelderland, South: provinces of Zeeland, Noord-Brabant and Limburg. HR: hazard ratio



Supplementary figure 2 Distribution of the propensity score in the CT-first patients and patients without a CT-first strategy before (solid) and after (dashed) propensity score matching.



Supplementary figure 3 Kaplan-Meier curves for cardiovascular mortality of the CT-first and the without a CT-first study population.

Chapter 4

Coronary calcification measures predict mortality in symptomatic women and men

Klaske R. Siegersma[¥], Floor Groepenhoff[¥], Anouk M. Eikendal, Willemijn J. op den Brouw, Tim Leiner, Yolande Appelman, Igor I. Tulevski, G. Aernout Somsen, N. Charlotte Onland-Moret, Leonard Hofstra[*], Hester M. den Ruijter[*]

[¥,*] These authors contributed equally

Submitted

ABSTRACT

Objective To assess the prognostic value of absolute and sex-, age-, and race/ethnicity-specific (MESA) percentiles of coronary artery calcification in symptomatic women and men.

Methods The study population consisted of 4985 symptomatic patients (2793 women, 56%) visiting a diagnostic outpatient cardiology clinic between 2009 and 2018 who were referred for cardiac computed tomography (CT) to determine the coronary artery calcification score (CACS). Regular care data was used and these data were linked to the databases of Statistics Netherlands for all-cause mortality data. Kaplan–Meier curves, multivariate Cox proportional hazards regression and concordance statistics were used to evaluate the prognostic value of CACS and MESA percentiles. Women were older compared to men (60 vs. 59 years).

Results Median CACS was 0 (IQR 0-54) in women and 42 (IQR 0-54) in men. After a median follow up of 4.4 years (IQR 3.1-6.3), 116 (2.3%; 53 women and 63 men) patients died. MESA percentiles did not perform better compared to absolute CACS (C-statistic 0.65, 95% confidence interval [CI] 0.57-0.73, vs 0.66, 95% CI 0.58-0.74, in women and 0.59, 95% CI 0.51-0.67, vs 0.62, 95% CI 0.55-0.69, in men, for the percentiles and absolute CACS, respectively).

Conclusion In symptomatic individuals absolute CACS predicts mortality with a moderately good performance. MESA percentiles did not perform better compared to absolute CACS, thus there is no need to use them. Including degree of stenosis in the model might slightly improved mortality risk prediction in women, but not in men.

INTRODUCTION

In asymptomatic women and men, calcification of the coronary arteries is proven to be a strong predictor for mortality.[1] In individuals with symptoms suspicious for cardiac disease, but without coronary artery calcification, the presence of obstructive coronary artery disease (CAD) is low and their long-term prognosis is good.[2] However, the value of coronary artery calcification in these symptomatic individuals is not well established.[3] Non-contrast enhanced computed tomography (CT) and coronary CT angiography are used to evaluate the calcification of the coronary arteries. Based on the results of this diagnostic assessment, the amount of coronary artery calcium is quantified by the coronary artery calcium score (CACS) according to Agatston.[4] CACS, as a continuous variable, is often categorized for clinical use.[5–7]

Several studies demonstrated sex differences in the amount and type of atherosclerotic plaques in symptomatic patients. Development of coronary artery calcification is on average delayed by 10 years in symptomatic women compared to men and onset of coronary artery calcification starts at an earlier age in men than in women.[5,8] Furthermore, symptomatic men mostly have calcified plaques while women predominantly have mixed or non-calcified plaques.[9,10] Therefore, use of absolute CACS to estimate mortality risk may lead to false reassurance in women with low CACS, as symptomatic women might have CAD caused by non-calcified plaques.[11] Thus, CACS may have a different prognostic value in symptomatic women compared to men[12,13], although literature is not consistent.[14]

The presence of coronary artery calcification also differs between ethnicities[15,16] and increases with age[5,6,8,17]. Therefore, the Multi-Ethnic Study of Atherosclerosis (MESA) reported sex-, age-, and race/ethnicity-specific percentiles for CACS in the general and asymptomatic population.[18,19] However, the question remains if these percentiles are a better discriminator of risk compared to absolute CACS in symptomatic patients in a real-world cardiology setting.

To address these issues, we first studied the prognostic value of coronary artery calcification measures in a symptomatic population in a sex-stratified manner, as CACS is mainly assessed and used as a risk marker in these symptomatic women and men.[20–22] Both absolute measures, as reflected by CACS, and MESA percentiles[18], were evaluated as measures to reflect the amount of calcification. Second, we evaluated whether degree of coronary stenosis at CT angiography increases the discriminative prognostic value when added to the model based on CACS.

METHODS

Patient selection

Individual patient data from electronic health records (CardioPortal™, Cardiology Centers of the Netherlands proprietary electronic health records, EHR) was retrieved from thirteen Dutch outpatient cardiology clinics (Cardiology Centers of the Netherlands, CCN) between 2007 and 2018. A detailed description of this database has been previously published.[23]. All included patients were symptomatic, i.e. they had cardiovascular complaints and were referred by the general practitioner to a cardiovascular screening center. We analyzed a selection of patients that underwent cardiac CT as part of clinical care, resulting in a study population of 4985 women and men, aged between 45 and 85 years (Figure 1). Standardized cardiovascular workup was performed and documented for these patients.[23] The Medical Research Ethics Committee of the University Medical Center Utrecht declared that research with the CCN database does not meet the Dutch Medical Research Involving Human Subjects Act (proposal number 17/359).

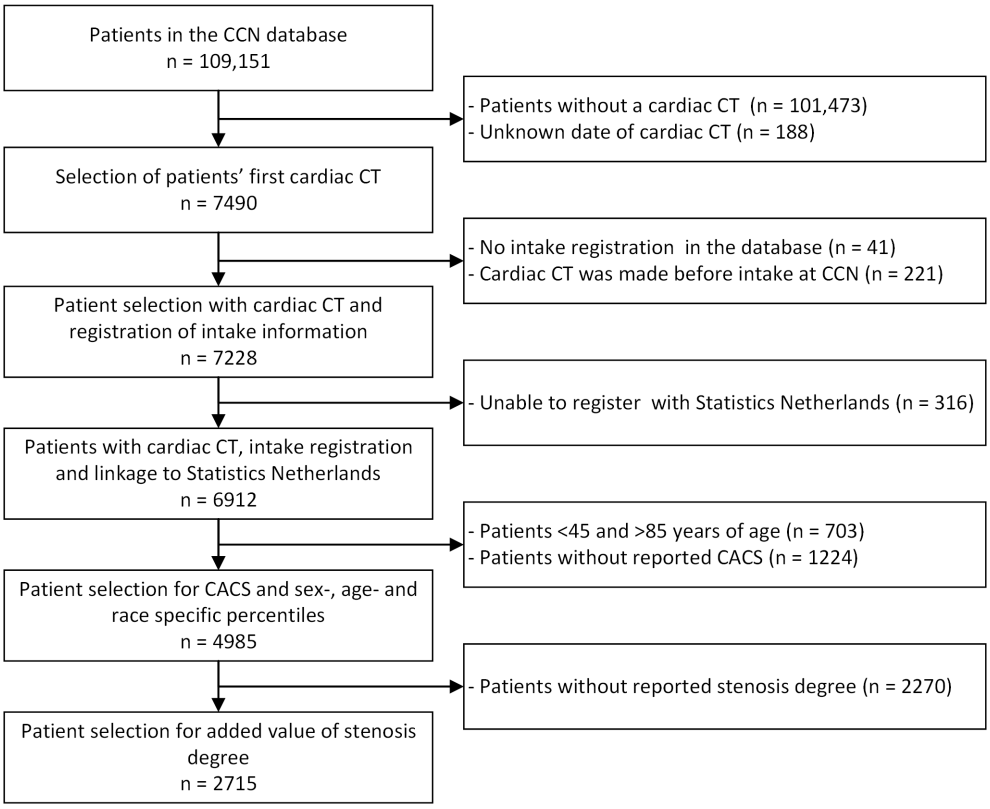


Figure 1 Patient selection for CACS and sex-, age- and race/ethnicity specific analysis and for analysis of addition of stenosis degree.

Calcium score and degree of stenosis assessment

The cardiac CT scanning protocol consisted of a non-contrast enhanced scan to evaluate CACS and a contrast-enhanced protocol for coronary angiography. Type of CT-scanner was determined by availability in the referred centers. Results of diagnostic imaging were reported in free text, which was transformed into different features, e.g. CACS and degree of stenosis in one of the main coronary arteries. Sex-, age-, and race/ethnicity-specific percentiles were based on retrieved absolute CACS using the previously reported percentile tables from the MESA.[19] As race/ethnicity was not structurally reported in the database, we used Caucasian percentiles in the main analysis.

As CACS is clinically used in different categories, we categorized CACS into the following groups: zero CACS, CACS 1 to 100, CACS 101 to 400 and CACS >400. For the MESA percentiles the following categories were used: no CAC, ≤75th percentile, 75th to 90th percentile, and >90th percentile. In 2715 (54%) of these patients, degree of stenosis determined by CT angiography was also documented in free text. A comparison was made for baseline characteristics of the complete CT population in which CACS was reported and the CT population in whom degree of stenosis was additionally documented. The available CT angiography results were classified into grades of stenosis, namely 0%, 1-24%, 25-49%, 50-70%, 71-99% and 100% stenosis.[24] Qualitative indications of degree of stenosis were discussed with two cardiologists to enable quantification of stenosis degree. This resulted in the following conversion from qualitative to quantative; for 1-24% *any, minimal, minor*, for 25-49% *partial, diffuse, mild, not-significant, non-significant*, for 50%-70% *important, clear, significant, intermediate*, for 71-99% *severe, high grade* and for 100% *occlusion*. As (high risk) plaque characteristics were irregularly and unstructured mentioned in free text, these features were not taken into account for analysis.

Outcome assessment

Information on patients' country of origin, mortality and cause of death was obtained by linkage to the population registry of Statistics Netherlands. Event rates are only exactly reported when 10 or more events were included, following regulations of Statistics Netherlands to avoid risk of personal disclosure. In all other cases, the number of events and percentages were reported as "<10" with the corresponding percentage.

Statistical analyses

The baseline characteristics of the datasets were described as mean +/- standard deviation (SD) or median with interquartile range (IQR) when appropriate. We estimated survival functions using Kaplan-Meier curves. Cox proportional hazards regression analysis was performed to study the predictive value of coronary artery calcification (absolute CACS as calculated by the Agatston score[4] and MESA percentiles[19]) and mortality. For the model based on the MESA percentiles, a binary variable was added to indicate whether

Table 1 Baseline characteristics of study population stratified by sex

	Overall	Women	Men
n	4985	2793	2192
Age in years (mean (SD))	59 (8)	60 (8)	59 (8)
Body mass index (mean (SD))	27 (5)	27 (5)	27 (4)
Originated from Europe (n, %)	4309 (86)	2411 (86)	1898 (87)
Complaints (n, %)			
Chest pain or discomfort	2683 (54)	1585 (57)	1098 (50)
Dyspnea	538 (11)	329 (12)	209 (10)
Fatigue	172 (4)	85 (3)	87 (4)
Palpitations	463 (9)	290 (10)	173 (8)
Collapse	27 (1)	<10	>10
Smoking status (n, %)			
Current	1476 (32)	859 (33)	617 (30)
Former	1650 (36)	875 (34)	775 (38)
Never	1504 (33)	854 (33)	650 (32)
Diabetes Mellitus (n, %)	396 (8)	213 (8)	183 (8)
Hypertension (n, %)	1581 (32)	954 (34)	627 (29)
Dyslipidemia (n, %)	850 (17)	477 (17)	373 (17)
CACS score (median [IQR])	8 [0-121]	0 [0-54]	42 [0-278]
CACS category (n, %)			
0	1956 (39)	1387 (50)	569 (26)
1-100	1673 (34)	899 (32)	774 (35)
101-400	776 (16)	340 (12)	436 (20)
>400	580 (12)	167 (6)	413 (19)
Examinations during follow-up (n, %)			
At least one CAG	812 (16)	309 (11)	503 (23)
At least one PCI or CABG	307 (6)	89 (3)	218 (10)
All-cause mortality (n, %)	116 (2)	53 (2)	63 (3)
Cardiovascular mortality (n, %)	22 (0.4)		
Years of follow up (median [IQR])	4.5 [3.1-6.3]	4.5 [3.1-6.3]	4.5 [3.1-6.3]

IQR: interquartile range, SD: standard deviation, CACS: coronary artery calcium score; CAC: coronary artery calcium; CAG: coronary angiography; PCI: percutaneous coronary intervention; CABG: coronary artery bypass graft.

CACS was positive (>0) at baseline. The addition of a variable to indicate the presence or absence of CACS is needed to correct for the possible discontinuity in the MESA percentiles between individuals with a CACS of zero or any positive continuous CACS. We first constructed models including CACS or MESA percentiles as continuous variables. Subsequently, we constructed models with the previously described categorized CACS and MESA percentiles. The concordance statistic (C-statistic) with 95% confidence intervals (CI) was used to assess the ability of these models to discriminate patients at risk for

mortality. A higher C-statistic indicates a better fit and higher prognostic power. Subsequently, we evaluated whether addition of degree of stenosis on top of absolute CACS improved the prognostic power of the Cox proportional hazards model. All analyses were stratified by sex. Statistical analyses were performed in R (version 4.0.2).

Sensitivity analyses

The race groups within the MESA percentiles as defined by the MESA19 were not transferable to patients' country of origin as documented in the population registry of Statistics Netherlands. In our primary analyses we calculated MESA percentiles based on the Caucasian race. In a sensitivity analysis we excluded all patients born outside Europe, using this individual's characteristic as a surrogate for a different race.

Patient and public involvement

Patients were not involved in any stage of this research process.

RESULTS

Baseline characteristics

Baseline characteristics of the selected population are displayed in Table 1. Of the 4985 patients, 2793 (56%) patients were women. On average, women were one year older compared to men (60 vs. 59). Other cardiovascular risk factors were similar between women and men. Both sexes primarily presented with chest pain or discomfort. Median CACS in women was 0 (IQR 0-54), and 42 (IQR 0-54) in men. CACS categories of 0, 1-100, 100-400 and > 400 were respectively present in 1387 (50%), 899 (32%), 340 (12%) and 167 (6%) women in this study. In men this distribution was 569 (26%), 774 (35%), 436 (20%) and 413 (19%), respectively. Higher CACS was seen in individuals with higher age and a higher prevalence of hypertension and dyslipidemia in both sexes (Supplementary table 1).

The association between coronary calcification and mortality

After a median follow up of 4 years (IQR 3-6), 116 (53 women and 63 men) patients died, of which 22 were attributed to cardiovascular mortality. Mortality rate was higher in individuals who had higher CACS (women: 1%, 2%, 4% and <6%, men: <2%, 3%, 3% and 5%, for, respectively CACS categories of 0, 1-100, 100-400 and > 400, Supplementary table 1). Figure 2 shows the survival over time per absolute CACS category for women and men. Survival over time for categorized MESA percentiles are shown in Figure 3. Overall, women had a better survival. Individuals with higher levels of absolute CACS showed lower survival rates compared to low CACS levels. This relation was also seen with higher MESA percentiles, albeit less strong. However, differences were small.

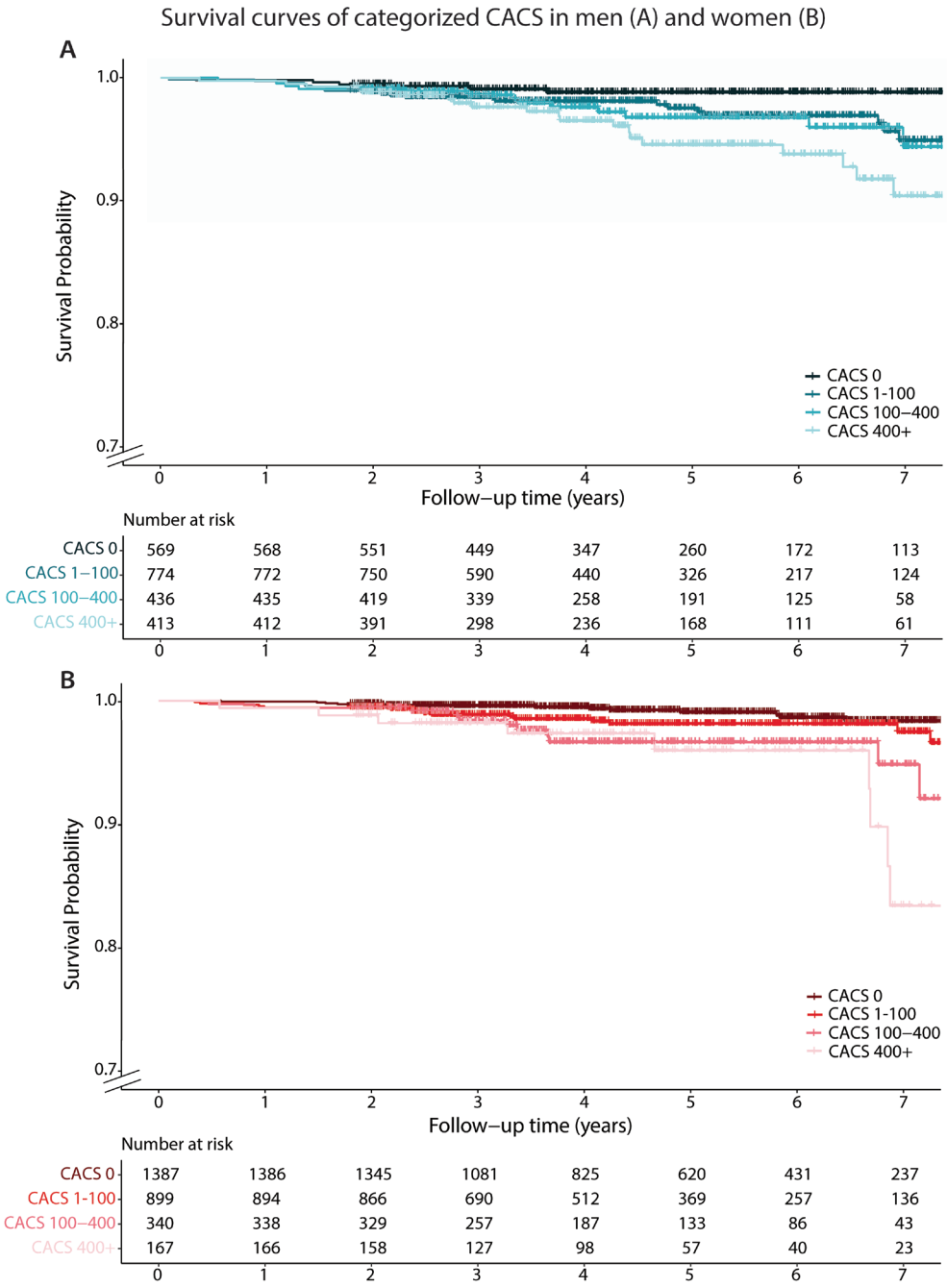


Figure 2 Survival of men (A) and women (B) by absolute CACS.

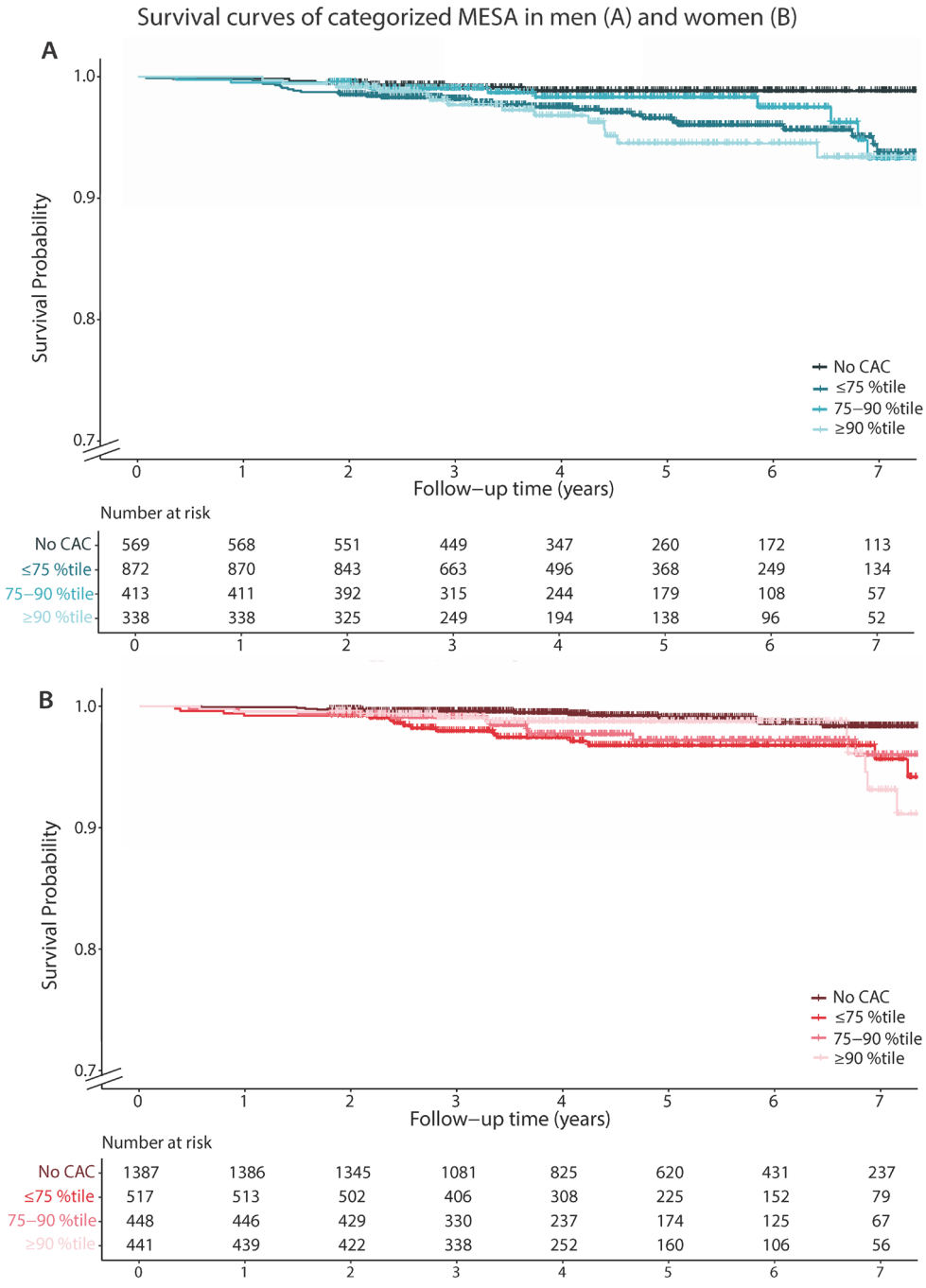


Figure 3 Survival of men (A) and women (B) by age-, sex- and race/ethnicity specific percentiles (MESA).

Table 2 Mortality prediction as a function of absolute and age-, sex-, and race/ethnicity specific percentiles of CACS in women (above) and men (below).

	Model	n	Events	HR (95% CI)	C-statistic (95% CI)
Women	**Absolute CACS**				
	Continuous	2793	53	1.0 (1.0-1.0)	0.66 (0.58-0.74)
	Categorized				0.65 (0.57-0.73)
	CACS 0	1387	14	reference	
	CACS 1-100	899	18	2.1 (1.0-2.1)	
	CACS 101-400	340	12	4.0 (1.8-8.6)	
	CACS >400	167	<10	5.7 (2.5-13.2)	
	Sex-, age-, and race/ethnicity-specific percentiles				
	Continuous adjusted for any calcification	2793	53		0.65 (0.57-0.73)
	MESA percentile			0.4 (0.1 - 1.5)	
	Any calcification			6.1 (1.9-19.1)	
	Categorized				0.64 (0.57-0.72)
	No calcification	1387	14	reference	
	< 75th percentile	517	18	3.5 (1.8-7.1)	
	75th-90th percentile	448	11	2.7 (1.2-5.9)	
	>90th percentile	441	10	2.5 (1.1-5.6)	
Men	**Absolute CACS**				
	Continuous	2192	63	1.0 (1.0-1.0)	0.62 (0.55-0.69)
	Categorized				0.60 (0.53-0.67)
	CACS 0	569	<10	reference	
	CACS 1-100	774	23	3.0 (1.2-7.5)	
	CACS 101-400	436	14	3.3 (1.3-8.7)	
	CACS >400	413	20	5.1 (2.0-12.7)	
	Sex-, age-, and race/ethnicity-specific percentiles				
	Continuous adjusted for any calcification	2192	63		0.59 (0.51-0.67)
	MESA percentile			0.5 (0.2-1.5)	
	Any calcification			5.7 (1.9-16.8)	
	Categorized				0.61 (0.54-0.67)
	No calcification	569	<10	reference	
	< 75th percentile	872	33	(1.6-9.3)	
	75th-90th percentile	413	10	2.5 (0.9-7.0)	
	>90th percentile	338	14	4.3 (1.6-11.1)	

CACS: coronary artery calcium score, HR: hazard ratio, C-statistic: concordance-statistic, CI: confidence interval

Mortality rates, hazards ratios and C-statistics are displayed in Table 2 for women and men. For continuous calcification measures, the discriminative ability of absolute CACS was moderate in both women and men (C-statistic 0.66, 95% CI 0.58-0.74 in women, and

0.62, 95% CI 0.55-0.69, in men). The discriminative ability of absolute CACS was similar to models based on MESA percentiles (0.65, 95% CI 0.57-0.73 in women, 0.59, 95% CI 0.51-0.67 in men). Results were similar for the models that used categorical classifications of CACS and MESA percentiles instead of continuous measures (CACS C-statistic 0.65, 95% CI 0.57-0.73 and 0.60, 95% CI 0.53-0.67, in women and men respectively, MESA percentiles C-statistic 0.64, 95% CI 0.57-0.73 and 0.61, 95% CI 0.54-0.67 in women and men respectively).

The sensitivity analysis in which we excluded all patients in whom the country of birth was documented to be outside Europe (n=676, 13.5%), as a surrogate for a different race, showed similar results. The European population comprised 2411 women and 1898 men. The C-statistic for absolute CACS was comparable to MESA percentiles (in women, 0.67, 95% CI 0.59-0.76, vs 0.66, 95% CI 0.58-0.75, and in men 0.64, 95% CI 0.55-0.72, vs 0.59, 95% CI 0.51-0.67). Results of the sensitivity analysis are displayed in Supplementary table 2.

Stenosis degree and the association between coronary calcification and mortality

Stenosis severity by CT angiography was documented for 1330 (60.7%) men and 1385 (49.6%) women. Baseline characteristics of these women and men are depicted in Table 3. Supplementary table 3 shows the baseline characteristics of the CT population in which CACS was reported and the CT population in whom degree of stenosis was additionally documented. During a median follow-up of 5 years (IQR 3-6 years), 46 (3.5%) men and 22 (1.5%) women died. Compared to the population in which only CACS was available, these women and men did not differ in baseline characteristics (for direct comparison and baseline characteristics per CACS category see Supplementary table 3 and 4, respectively). In the population in whom both CACS and information on stenosis degree was available, the relation between calcification measures and mortality was comparable to the relation found in the total population (Table 4). In men, the discriminative power of the model did not improve when degree of stenosis was added (C-statistic changed from 0.63, 95% CI 0.55-0.71 based on CACS to 0.59, 95% CI 0.51-0.67 after addition of stenosis to the model). In women, the performance to predict mortality improved slightly (C-statistic 0.68, 95% CI 0.58-0.78, and 0.72, 95% CI 0.61-0.83, respectively). However, this improvement was not significant and thus no hard conclusions can be drawn.

DISCUSSION

Our data showed that absolute CACS and MESA percentiles perform equally well in predicting mortality in symptomatic women and men who visit outpatient cardiology clinics in a real-world setting. Hence, for discrimination of mortality in symptomatic individuals there is no need for MESA percentiles to quantify coronary artery calcification. Absolute CACS predicts mortality with moderate performance, comparable to performance in as-

Table 3 Baseline characteristics of women and men that underwent cardiac CT angiography.

	Overall	Women	Men
n	2715	1385	1330
Age in years (mean (SD))	60 (8)	60 (8)	59 (8)
Body mass index (mean (SD))	27 (4)	27 (5)	27 (4)
Complaints (n, %)			
Chest pain or discomfort	1499 (55)	801 (58)	698 (53)
Dyspnea	295 (11)	167 (12)	128 (10)
Fatigue	93 (3)	47 (3)	46 (4)
Palpitations	272 (10)	163 (12)	109 (8)
Collapse	18 (1)	<10	>10
Smoking status (n, %)			
Current	809 (32)	428 (34)	381 (31)
Former	936 (37)	461 (36)	475 (38)
Never	765 (31)	380 (30)	385 (31)
Diabetes Mellitus (n, %)	211 (8)	97 (7)	114 (9)
Hypertension (n, %)	885 (33)	518 (38)	367 (28)
Dyslipidemia (n, %)	497 (18)	263 (19)	234 (18)
CACS score (median [IQR])	34 [0-162]	14 [0-92]	67 [6-267]
CACS category (n, %)			
0	682 (25)	472 (34)	210 (16)
1-100	1150 (42)	585 (42)	565 (43)
101-400	598 (22)	251 (18)	347 (26)
>400	285 (11)	77 (6)	208 (16)
Examinations during follow-up (n, %)			
At least one CAG	553 (20)	205 (15)	348 (26)
At least one PCI or CABG	234 (9)	66 (5)	168 (13)
All-cause mortality (n, %)	68 (3)	22 (2)	46 (4)
Cardiovascular mortality (n, %)	10 (0.4)		
Years of follow up (median [IQR])	4.6 [3.2-6.3]	4.6 [3.2-6.3]	4.7 [3.2-6.4]

IQR: interquartile range, SD: standard deviation, CACS: coronary artery calcium score; CAC: coronary artery calcium; CAG: coronary angiography; PCI: percutaneous coronary intervention; CABG: coronary artery bypass graft.

ymptomatic inidivuals.[19] Finally, the data hint that in women, the discriminative power of CACS for mortality might be higher when degree of stenosis was included in the model. However, these results should be interpreted with caution as improvement was not significant. In men, addition of degree of stenosis did not result in better prediction of mortality. This subtle sex-difference might be due to the presence of non-calcified plaques causing symptoms in women.[9] This type of plaques remains (partly) unappreciated when using CACS only for mortality prediction.

When comparing the results of the result of the prognostic value of CACS to other pub-

Table 4 Mortality prediction as a function of absolute CACS and degree of stenosis in women (above) and men (below).

		Model	n	Events	HR (95% CI)	C-statistic (95% CI)
Women	**Absolute CACS**					
		Continuous	1385	22	1.0 (1.0-1.0)	0.68 (0.58-0.78)
		Degree of stenosis				0.72 (0.61-0.83)
		0%	527	<10	reference	
	+ Stenosis	1%-49%	563	14	6.9 (1.6-30.6)	
		50%-70%	168	<10	4.9 (0.9-28.6)	
		>70%	127	<10	2.8 (0.3-22.4)	
Men	**Absolute CACS**					
		Continuous	1330	46	1.0 (1.0-1.0)	0.63 (0.55-0.71)
		Degree of stenosis				0.72 (0.61-0.83)
		0%	265	<10	reference	0.59 (0.51-0.67)
	+ Stenosis	1%-49%	551	22	4.5 (1.3-15.3)	
		50%-70%	263	10	3.9 (1.1-14.7)	
		>70%	251	11	4.2 (1.1-16.3)	

CACS: coronary artery calcium score, HR: hazard ratio, CI: confidence interval

lications, a similar result was presented by Engbers et al.[14] in which they sought to evaluate gender-specific (n=3705, 61% women) prognostic value of CACS on top of SPECT myocardial perfusion imaging. No gender-specific differences were found and the hazard ratios described in this study are similar to the hazard ratios for the CACS categories in both sexes in our study. Another study that focused on symptomatic individuals (n=3840, 51% women) suspected of CAD showed that a high prognostic value of CACS, which further increased after addition of stenosis degree.[3] More variables, i.e. degree of stenosis, specific (high risk) plaque characteristics, were incorporated in prognostic models. However, their analysis was not stratified by sex, hampering any comparison to our data and impeding analysis of sex differences. They also showed that the recently developed Coronary Artery Disease Reporting and Data System (CAD-RADS)[24] classification provides the highest prognostic value for cardiovascular events.[3]

Our results in symptomatic women and men are in line with results of the MESA, which was conducted in non-symptomatic individuals from the general population. The MESA demonstrated that absolute CACS outperforms MESA percentiles for event prediction.[19] We also found that the discriminative capacity remains intact when categorizing CACS. This finding is valuable for clinical use, as hazard ratios derived from categories are easier to interpret than continuous values. Furthermore, studies in asymptomatic individuals concluded that CACS predicted mortality risk equally well for both sexes[1,12], despite the findings that women had lower CACS compared to men[1] and that significant sex differences were present in cumulative mortality[12]. Most studies reported that CACS may be

a better predictor for mortality in women compared to men.[12,21,22] This suggestion is in line with the slightly higher C-statistic for the CACS-model we described in women. In addition, sex-specific CACS percentiles tend to better stratify risk in women than men as opposed to absolute scores.[21] The reason for discrepancy between results might be the inclusion of an asymptomatic older population in the study by Wang et al.[21]

Strengths and limitations

A strength of the presented study is the use of a real-world population in whom cardiac CT is often used as a primary diagnostic, yet its power is not often studied. This population is best described as a symptomatic population, referred to a specialized cardiac screening center, which is positioned between general practitioners and hospital care. Furthermore, linkage to the database of Statistics Netherlands provided long-term follow-up data on mortality and information on country of birth.

Use of EHR data also has inherent limitations. First, data were not primarily collected for study purposes. Therefore, part of the population that underwent cardiac CT could not be included, due to insufficient documentation of CT results. Second, as CT angiography was only performed based on referral by the cardiologist and/or local scan protocol, not all patients underwent CT angiography. Thus, the subanalysis focusing on degree of stenosis, based on CT angiography results, was only performed in a subset of patients. This significantly reduced the power of our statistical analysis and might have led to selection bias. To evaluate this bias, we compared the overall and sub-population and repeated the analyses in the subpopulation. The populations were comparable on known baseline characteristics. Moreover, the results of the use of CACS and MESA percentiles for risk stratification did not significantly differ. This suggests that increased discriminative power of CACS when degree of stenosis was added to the model is generalizable. Third, as stenosis severity was retrieved from text reports in which degree of stenosis was not always quantified, the interpretation of the grade of stenosis could lead to uncertainty in our data. Fourth, we did not take treatment differences into account between different groups. These differences may explain the decreasing hazard ratio with an increasing degree of coronary stenosis, although this decrease was not significant. Finally, as we have used country of origin as a surrogate for race/ethnicity, the categories used in the original MESA calculations and our study population were not identical. To assess whether this has influenced our results we performed a sensitivity analysis in which we included women and men who originated from Europe, assuming they all have the Caucasian race. These results were not significantly different from the main analyses. Moreover, as our population was primarily coming from Europe, our results might not be applicable to individuals of other ethnic populations.

We found that adding degree of stenosis to the survival model slightly improved the

discriminative power in women, but not in men. However, this improvement in women was not statistically significant and confidence intervals largely overlapped. Even though we showed a moderate prognostic value of CACS in both symptomatic women and men, our results warrant evaluation beyond CACS alone for optimal risk prediction as C-statistics were below 0.70. This might be especially relevant in women, in whom information on the degree of stenosis potentially has added value due to presence of non-calcified plaques. We were unable to assess the added value of degree of stenosis properly due to incomplete data. Nevertheless, the difference between women and men we found after degree of stenosis was added to the model stresses the importance of a sex-specific view on CAD. Replication of this study in larger trials or populations is going to contribute to confirm these hypotheses, regarding the importance of non-calcified plaques in women for mortality risk.

CONCLUSION

In symptomatic individuals absolute CACS predicts mortality with a moderately good performance. MESA percentiles did not perform better compared to absolute CACS, thus there is no need to use them. Including degree of stenosis in the model might slightly improved mortality risk prediction in women, but not in men.

REFERENCES

1. Nakanishi R, Li D, Blaha MJ, et al. All-cause mortality by age and gender based on coronary artery calcium scores. Eur Heart *J Cardiovasc Imaging*. 2016;17(11):1305-1314. doi:10.1093/ehjci/jev328
2. Mittal TK, Pottle A, Nicol E, et al. Prevalence of obstructive coronary artery disease and prognosis in patients with stable symptoms and a zero-coronary calcium score. *Eur Heart J Cardiovasc Imaging*. 2017;18(8):922-929. doi:10.1093/ehjci/jex037
3. Bittner DO, Mayrhofer T, Budoff M, et al. Prognostic Value of Coronary CTA in Stable Chest Pain: CAD-RADS, CAC, and Cardiovascular Events in PROMISE. *JACC Cardiovasc Imaging*. 2020;13(7):1534-1545. doi:10.1016/j.jcmg.2019.09.012
4. Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte M, Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. *J Am Coll Cardiol*. 1990;15(4):827-832. doi:10.1016/0735-1097(90)90282-T
5. Nicoll R, Wiklund U, Zhao Y, et al. Gender and age effects on risk factor-based prediction of coronary artery calcium in symptomatic patients: A Euro-CCAD study. *Atherosclerosis*. 2016;252:32-39. doi:10.1016/j.atherosclerosis.2016.07.906
6. Raggi P, Gongora MC, Gopal A, Callister TQ, Budoff M, Shaw LJ. Coronary Artery Calcium to Predict All-Cause Mortality in Elderly Men and Women. *J Am Coll Cardiol*. 2008;52(1):17-23. doi:10.1016/j.jacc.2008.04.004
7. Rumberger JA, Brundage BH, Rader DJ, Kondos G. Electron Beam CT CAC scanning-a review and guidelines for use in asymptomatic persons. *Mayo Clin Proc*. 1999;74:243-252.
8. Kim BS, Chan N, Hsu G, et al. Sex Differences in Coronary Arterial Calcification in Symptomatic Patients. *Am J Cardiol*. 2021;149(January 2016):16-20. doi:10.1016/j.amjcard.2021.03.025
9. Plank F, Beyer C, Friedrich G, Wildauer M, Feuchtner G. Sex differences in coronary artery plaque composition detected by coronary computed tomography: quantitative and qualitative analysis. *Netherlands Hear J*. 2019;27(5):272-280. doi:10.1007/s12471-019-1234-5
10. Williams MC, Kwiecinski J, Doris M, et al. Sex-Specific Computed Tomography Coronary Plaque Characterization and Risk of Myocardial Infarction. *JACC Cardiovasc Imaging*. Published online 2021. doi:10.1016/j.jcmg.2021.03.004
11. Hussain A, Ballantyne CM, Nambi V. Zero Coronary Artery Calcium Score: Desirable, but Enough? *Circulation*. Published online 2020:917-919. doi:10.1161/CIRCULATIONAHA.119.045026
12. Raggi P, Shaw LJ, Berman DS, Callister TQ. Gender-based differences in the prognostic value of coronary calcification. *J Women's Heal*. 2004;13(3):273-283. doi:10.1089/154099904323016437
13. Bellasi A, Lacey C, Taylor AJ, et al. Comparison of Prognostic Usefulness of Coronary Artery Calcium in Men Versus Women (Results from a Meta- and Pooled Analysis Estimating All-Cause Mortality and Coronary Heart Disease Death or Myocardial Infarction). *Am J Cardiol*. 2007;100(3):409-414. doi:10.1016/j.amjcard.2007.03.037
14. Engbers EM, Timmer JR, Ottervanger JP, Mouden M, Knollema S, Jager PL. Impact of Gender on the Prognostic Value of Coronary Artery Calcium in Symptomatic Patients With Normal Single-Photon Emission Computed Tomography Myocardial Perfusion. *Am J Cardiol*. 2016;118(11):1611-1615. doi:10.1016/j.amjcard.2016.08.037
15. Bild DE, Detrano R, Peterson D, et al. Ethnic differences in coronary calcification: The Multi-Ethnic Study of Atherosclerosis (MESA). *Circulation*. 2005;111(10):1313-1320. doi:10.1161/01.

CIR.0000157730.94423.4B
16. Nasir K, Shaw LJ, Liu ST, et al. Ethnic Differences in the Prognostic Value of Coronary Artery Calcification for All-Cause Mortality. *J Am Coll Cardiol*. 2007;50(10):953-960. doi:10.1016/j.jacc.2007.03.066
17. Hoff JA, Chomka E V., Krainik AJ, Daviglus M, Rich S, Kondos GT. Age and gender distributions of coronary artery calcium detected by electron beam tomography in 35,246 adults. *Am J Cardiol.* 2001;87(12):1335-1339. doi:10.1016/S0002-9149(01)01548-X
18. Budoff MJ, Nasir K, McClelland RL, et al. Coronary Calcium Predicts Events Better With Absolute Calcium Scores Than Age-Sex-Race/Ethnicity Percentiles. MESA (Multi-Ethnic Study of Atherosclerosis). *J Am Coll Cardiol*. 2009;53(4):345-352. doi:10.1016/j.jacc.2008.07.072
19. McClelland RL, Chung H, Detrano R, Post W, Kronmal RA. Distribution of coronary artery calcium by race, gender, and age: Results from the Multi-Ethnic Study of Atherosclerosis (MESA). *Circulation*. 2006;113(1):30-37. doi:10.1161/CIRCULATIONAHA.105.580696
20. Shaw LJ, Min JK, Nasir K, et al. Sex differences in calcified plaque and long-term cardiovascular mortality: Observations from the CAC Consortium. *Eur Heart J*. 2018;39(41):3727-3735. doi:10.1093/eurheartj/ehy534
21. Wang FM, Rozanski A, Arnson Y, et al. Cardiovascular and All-Cause Mortality Risk by Coronary Artery Calcium Scores and Percentiles Among Older Adult Males and Females. *Am J Med.* 2021;134(3):341-350.e1. doi:10.1016/j.amjmed.2020.07.024
22. Kelkar AA, Schultz WM, Khosa F, et al. Long-Term Prognosis after Coronary Artery Calcium Scoring among Low-Intermediate Risk Women and Men. *Circ Cardiovasc Imaging*. 2016;9(4):1-7. doi:10.1161/CIRCIMAGING.115.003742
23. Bots SH, Siegersma KR, Onland-Moret NC, et al. Routine clinical care data from thirteen cardiac outpatient clinics: design of the Cardiology Centers of the Netherlands (CCN) database. *BMC Cardiovasc Disord*. 2021;21(1):1-9. doi:10.1186/s12872-021-02020-7
24. Foldyna B, Szilveszter B, Scholtz JE, Banerji D, Maurovich-Horvat P, Hoffmann U. CAD-RADS – a new clinical decision support tool for coronary computed tomography angiography. *Eur Radiol*. 2018;28(4):1365-1372. doi:10.1007/s00330-017-5105-4

SUPPLEMENTARY MATERIALS

Supplementary table 1 Baseline characteristics of study population stratified by sex and CACS.

		CACS 0	CACS 1-100	CACS 101-400	CACS >400
Women	n	1387	899	340	167
	Age in years (mean (SD))	56 (7)	62 (8)	64 (8)	67 (8)
	Body mass index (mean (SD))	26 (5)	27 (5)	27 (5)	27 (5)
	Smoking status (n, %)				
	Current	441 (34)	270 (32)	99 (32)	49 (31)
	Former	383 (30)	301 (36)	129 (42)	62 (40)
	Never	465 (36)	262 (32)	82 (27)	45 (29)
	Diabetes Mellitus (n, %)	78 (6)	75 (8)	30 (9)	30 (18)
	Hypertension (n, %)	349 (25)	352 (39)	164 (48)	89 (53)
	Dyslipidemia (n, %)	167 (12)	179 (20)	82 (24)	49 (29)
	Degree of stenosis (n, %)				
	0%	432 (92)	82 (14)	11 (4)	<10
	1%-49%	34 (7)	399 (68)	110 (44)	>10
	50%-70%	<10	65 (11)	74 (30)	23 (30)
	>70%	<10	39 (7)	56 (22)	32 (42)
	Examinations during follow-up (n, %)				
	At least one CAG	>10	90 (10)	114 (34)	90 (54)
	At least one PCI or CABG	<10	24 (3)	23 (7)	41 (25)
	All-cause mortality (n, %)	14 (1)	18 (2)	12 (4)	<10
	Years of follow up (median [IQR])	4.6 [3.2-6.5]	4.4 [3.1-6.3]	4.3 [3.0-6.0]	4.3 [3.1-5.9]
Men	n	569	774	436	413
	Age in years (mean (SD))	54 (7)	58 (8)	61 (8)	64 (7)
	Body mass index (mean (SD))	27 (4)	27 (4)	27 (4)	27 (4)
	Smoking status (n, %)				
	Current	165 (32)	217 (30)	113 (28)	122 (32)
	Former	168 (32)	259 (36)	174 (42)	174 (45)
	Never	191 (37)	245 (34)	123 (30)	91 (24)
	Diabetes Mellitus, n (%)	23 (4)	64 (8)	35 (8)	61 (15)
	Hypertension, n (%)	113 (20)	197 (26)	153 (35)	164 (40)
	Dyslipidemia, n (%)	57 (10)	129 (17)	85 (20)	102 (25)
	Degree of Stenosis, n (%)				
	0%	181 (86.2)	68 (12)	14 (4)	<10
	1%-49%	25 (12)	355 (63)	128 (37)	40-45
	50%-70%	<10	94 (17)	109 (31)	57 (27)
	>70%	<10	48 (9)	96 (28)	106 (51)
	Examinations during follow-up, n (%)				
	At least one CAG	<10	109 (14)	147 (34)	239 (58)
	At least one PCI or CABG	<10	38 (5)	67 (15)	111 (27)
	All-cause mortality, n (%)	<10	23 (3)	14 (3)	20 (5)
	Years of follow up (median [IQR])	5 [3-6]	4 [3-6.]	4 [3-6]	4 [3-6]

IQR: interquartile range, SD: standard deviation, CACS: coronary artery calcium score; CAC: coronary artery calcium; CAG: coronary angiography; PCI: percutaneous coronary intervention; CABG: coronary artery bypass graft.

Supplementary table 2 Mortality prediction as a function of absolute and age-, sex- and ethnicity specific percentiles of CACS in women (above) and men (below), only performed in European women and men.

	Model	n	Events	HR (95% CI)	C-statistic (95% CI)
Women	**Absolute CACS**				
	Continuous	2411	47	1.0 (1.0-1.0)	0.67 (0.59-0.76)
	Categorized				0.67 (0.59-0.75)
	CACS 0	1185	11	reference	
	CACS 1-100	773	16	2.9 (1.1-5.2)	
	CACS 101-400	304	12	5.0 (2.2-11.3)	
	CACS >400	149	<10	6.3 (2.5-15.7)	
	Sex-, age-, and race/ethnicity-specific percentiles				
	Continuous adjusted for any calcification	2411	47		0.61 (0.58-0.75)
	MESA percentile			0.4 (0.1-1.7)	
	Any calcification			6.9 (2.0-23.7)	
	Categorized				0.66 (0.58-0.74)
	No calcification	1185	11	reference	
	< 75th percentile	466	16	3.8 (1.8-8.3)	
	75th-90th percentile	386	11	3.4 (1.5-8.0)	
	>90th percentile	374	<10	3.0 (1.2-7.1)	
Men	**Absolute CACS**				
	Continuous	1898	57	1.0 (1.0-1.0)	0.64 (0.56-0.72)
	Categorized				0.62 (0.55-0.70)
	CACS 0	480	<10	reference	
	CACS 1-100	665	20	3.9 (1.3-11.3)	
	CACS 101-400	380	13	4.4 (1.4-13.6)	
	CACS >400	373	20	7.1 (2.4-20.7)	
	Sex-, age-, and race/ethnicity-specific percentiles				
	Continuous adjusted for any calcification	1898	57		0.59 (0.51-0.67)
	MESA percentile			0.5 (0.2-1.7)	
	Any calcification			7.3 (2.1-25.6)	
	Categorized				0.62 (0.55-0.69)
	No calcification	480	<10	reference	
	< 75th percentile	763	30	5.1 (1.8-14.5)	
	75th-90th percentile	360	<10	3.3 (1.0-10.6)	
	>90th percentile	295	14	6.1 (2.0-18.6)	

CACS: coronary artery calcium score, HR: hazard ratio, CI: confidence interval

Supplementary table 3 Direct comparison of CT and CTA population

	CT population	Subgroup with CTA
n	4985	2715
Female (n, %)	2793 (56)	1385 (51)
Age in years (mean (SD))	59 (8)	59.66 (8)
Body mass index (mean (SD))	27 (4)	26.68 (4)
Complaints (n, %)		
Chest pain or discomfort	2683 (54)	1499 (55)
Dyspnea	538 (11)	295 (11)
Fatigue	172 (4)	93 (3)
Palpitations	463 (9)	272 (10)
Collapse	27 (0.5)	18 (0.7)
Smoking status (n, %)		
Current	1476 (32)	809 (32)
Former	1650 (36)	936 (37)
Never	1504 (33)	765 (31)
Diabetes Mellitus (n, %)	396 (8)	211 (8)
Hypertension (n, %)	1581 (32)	885 (33)
Dyslipidemia (n, %)	850 (17)	497 (18)
CACS score (median [IQR])	7.60 [0, 121]	34.00 [0, 162]
CACS category (n, %)		
0	1956 (39)	682 (25)
1-100	1673 (34)	1150 (42)
101-400	776 (16)	598 (22)
>400	580 (12)	285 (11)
All-cause mortality (n, %)	116 (2)	68 (3)
Cardiovascular mortality (n, %)	22 (0.4)	10 (0.4)

SD: standard deviation, IQR: interquartile range, CACS: coronary artery calcium score

Supplementary table 4 Baseline characteristics of CTA study population stratified by sex and CACS.

		CACS 0	CACS 1-100	CACS 101-400	CACS >400
Women	n	472	585	251	77
	Age in years (mean (SD))	55.90 (6.98)	61.32 (7.68)	64.49 (7.70)	66.45 (7.99)
	Body mass index (mean (SD))	26.22 (4.82)	26.71 (4.90)	27.19 (5.24)	26.90 (4.31)
	Smoking status (n, %)				
	Current	168 (39)	173 (32)	68 (29)	19 (27)
	Former	136 (32)	200 (37)	99 (43)	26 (37)
	Never	123 (29)	166 (31)	66 (28)	25 (36)
	Diabetes Mellitus (n, %)	22 (5)	43 (7)	21 (8)	10-15 (13-19)
	Hypertension (n, %)	135 (29)	221 (38)	122 (49)	40 (52)
	Dyslipidemia (n, %)	64 (14)	110 (19)	66 (26)	23 (30)
	Degree of Stenosis (n, %)				
	0%	432 (92)	82 (14)	11 (4)	<10 (<13)
	1%-49%	34 (7)	399 (68)	110 (44)	20-25 (25-32)
	50%-70%	<10 (<2)	65 (11)	74 (30)	23 (30)
	>70%	<10 (<2)	39 (7)	56 (22)	32 (42)
	Examinations during follow-up (n, %)				
	At least one CAG	<10 (<2)	72 (12)	96 (38)	32 (42)
	At least one PCI or CABG	<10 (<2)	23 (4)	22 (9)	20 (26)
	All-cause mortality (n, %)	<10 (<2)	<10 (<2)	<10 (<4)	<10 (<13)
	Years of follow up (median [IQR])	5 [3-7]	4 [3-6]	4 [3-6]	4 [3-5]
Men	n	210	565	347	208
	Age in years (mean (SD))	53.71 (6.40)	58.13 (7.69)	60.98 (8.05)	63.19 (7.55)
	Body mass index (mean (SD))	26.49 (4.03)	26.83 (3.67)	26.79 (3.79)	26.51 (3.42)
	Smoking status (n, %)				
	Current	75 (40)	160 (30)	85 (26)	61 (32)
	Former	55 (29)	193 (36)	146 (45)	81 (42)
	Never	57 (31)	181 (34)	96 (29)	51 (26)
	Diabetes Mellitus (n, %)	<10	48 (9)	28 (8)	29 (14)
	Hypertension (n, %)	34 (16)	143 (25)	118 (34)	72 (35)
	Dyslipidemia (n, %)	23 (11)	94 (17)	65 (19)	52 (25)
	Degree of Stenosis (n, %)				
	0%	181 (86)	68 (12)	14 (4)	<10 (<4)
	1%-49%	25 (12)	355 (63)	128 (37)	40-45 (19-22)
	50%-70%	<10 (<4)	94 (17)	109 (31)	57 (27)
	>70%	<10 (<4)	48 (9)	96 (28)	106 (51)
	Examinations during follow-up (n, %)				
	At least one CAG	<10 (<4)	92 (16)	127 (37)	123 (59)
	At least one PCI or CABG	<10 (<4)	33 (6)	61 (18)	73 (35)
	All-cause mortality (n, %)	<10 (<4)	21 (4)	12 (4)	12 (6)
	Years of follow up (median [IQR])	5 [4-7]	4 [3-6]	5 [3-6]	4 [3-6]

IQR: interquartile range, SD: standard deviation, CACS: coronary artery calcium score; CAC: coronary artery calcium; CAG: coronary angiography; PCI: percutaneous coronary intervention; CABG: coronary artery bypass graft.

Chapter 5A

NYHA class is strongly associated with mortality beyond heart failure in symptomatic women

Klaske R. Siegersma[¥], Floor Groepenhoff[¥], N. Charlotte Onland-Moret , Igor I. Tulevski, Leonard Hofstra, G. Aernout Somsen, Hester M. den Ruijter

[¥] These authors contributed equally

EHJ - Quality of Care and Clinical Outcomes 2021; 7(2): 214-215
(Published in short)

Cardiovascular disease in women remains the leading cause of mortality worldwide, which stresses the need for accurate risk prediction in women.[1] The New York Heart Association (NYHA) functional classification of patients with cardiovascular disease is a commonly used clinical scale to estimate the general condition of a patient.[2] It has been specifically designed and used for heart failure patients and has shown to be a valid and easy obtainable measure to predict mortality in these patients.[3] Nowadays, use of this clinical scale is not solely limited to heart failure patients but widely used by cardiologists to grade severity of several complaints, i.e. chest pain, dyspnoea and fatigue, and classify the condition of an individual.[4] However, most validation studies have been performed in men with acute heart failure and recent studies in more heterogeneous populations showed that the predictive value of NYHA class for mortality might differ between the sexes and patient domains.[3,5]

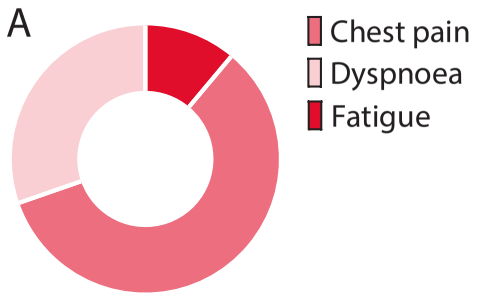
Due to subacute presentation of disease in women, an increased number of women are being evaluated at outpatient cardiology clinics over time. As the diagnosis is challenging, a variety of clinical scales are used and tests are performed to assess cardiovascular disease (risk). Yet, clinicians are in the dark which of these performs best in women, because originally diagnostic risk stratification tools are mainly developed in men. Since the NYHA classification scale is an easily obtainable non-invasive tool without any burden on the patient, it is often used in women at outpatient clinics. However, the value of NYHA class for prediction of mortality in this large and heterogeneous female population is lacking. Therefore, in women with a variety of symptoms, i.e. chest pain, dyspnoea and fatigue, we specifically studied the association between NYHA class and mortality in a large population of patients presenting at outpatient cardiology clinics.

We extracted electronic health record data of thirteen outpatient clinics (Cardiology Centers of the Netherlands) for individuals that have visited between 2007 and 2018 with a documented NYHA class for chest pain, dyspnoea or fatigue at their initial visit. Patients received a diagnostic work-up, including NYHA class for chest pain, dyspnoea or fatigue, blood tests, echocardiography, a stress and rest electrocardiogram, full anamnesis by a specialized nurse and a cardiologist' consult. Follow-up for mortality was performed by linkage to the population registry of Statistics Netherlands (CBS). We estimated survival functions using Kaplan-Meier method. Cox proportional hazards regression analysis was used to study the association between NYHA class and mortality corrected for age. All analyses were performed per sex and stratified by primary complaint (i.e. chest pain, dyspnoea and fatigue). All statistical analyses were performed in R (version 4.0.2). The Medical Research Ethics Committee of the UMCU waived the necessity for informed consent because research within the Cardiology Centers of the Netherlands database does not fall under the Dutch Medical Research Involving Human Subjects Act (Wet medisch-wetenschappelijk onderzoek met mensen, WMO).

Of the 9011 patients, 4782 (53%) were female of whom 1450 presented with dyspnoea, 2801 with chest pain and 531 with fatigue as primary complaint. NYHA class I, II and III-IV out of IV were respectively documented in 2196 (46%), 2077 (43%) and 509 (11%) women (for male patients, this distribution was 2114 (50%), 1688 (40%) and 428 (10%), respectively). Higher NYHA class at baseline was related to higher age. Other standard cardiovascular risk factors were similar between the NYHA classes (Figure 1, panel B). After eight years of follow up 354 (7%) women and 415 (10%) men died, of which 134 (38%) and 150 (36%), respectively, were classified as cardiovascular death.

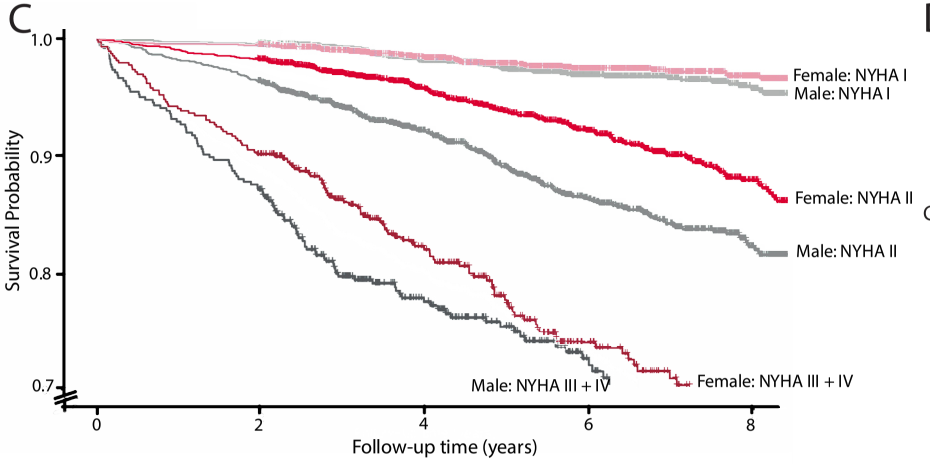
In all women, regardless of primary complaint, survival analysis showed that increased NYHA class was positively associated with mortality, both for all-cause (Figure 1, panel C) and cardiovascular (data not shown) mortality. Multivariable analysis corrected for age confirmed that women with NYHA class III-IV have a higher risk of mortality as compared with women presenting with NYHA class I (all women: hazard ratio [HR] 3.9, 95% confidence interval [CI] 2.8-5.5, chest pain: HR 2.4, 95% CI 1.3-4.6, dyspnoea: HR 2.6, 95% CI 1.5-4.6, fatigue: HR 2.5, 95% CI 1.0-6.0). Women suffering from complaints classified as NYHA class II showed a similar, but less evident, trend (all women: HR 1.7, 95% CI 1.3-2.3, chest pain: HR 1.4, 95% CI 0.8-2.2, dyspnoea: HR 1.2 95% CI 0.7-2.1, fatigue: HR 0.9, 95% CI 0.4-2.1). In all men, and in men presenting with dyspnoea, chest pain or fatigue separately, a similar relationship between NYHA class and (cardiovascular) mortality was found (NYHA class III-IV vs. I in all men: HR 4.4, 95% CI 3.2-6.0, chest pain: HR 2.3, 95% CI 1.4-3.8, dyspnoea: HR 4.1, 95% CI 2.4-7.8, fatigue: HR 7.9, 95% CI 2.0-31.3, data on cardiovascular mortality not shown). The association between NYHA class III-IV and mortality remained significant for both women and men after adjustment for SCORE (women: HR 7.8, 95% CI 4.9-12.2, men: HR 7.1, 95% CI 4.8-10.5).

Similar to men, in women presenting with chest pain, fatigue or dyspnoea at outpatient cardiology clinics functional grading of their complaints using NYHA classification provides important information on their (cardiovascular) mortality risk. As NYHA class is an easily obtainable, non-invasive measure, treating physicians are able to quickly get an idea of the physical status and survival risk of the visiting female patient. We have shown that this is the case for patients with a wide range of complaints and not solely for heart failure patients. Therefore, women suffering from NYHA class III-IV complaints may warrant a more aggressive diagnostic and therapeutic workup.



B

	NYHA Class I	NYHA Class II	NYHA Class III+IV
Female patients (n)	2196	2077	509
Mean age (SD)	55.6 (13.3)	63.7 (12.8)	69.8 (12.4)
Mean BMI (SD)	25.7 (4.7)	27.5 (5.2)	28.4 (6.4)
Current smoker (%)	782 (39.1)	758 (39.2)	182 (38.6)
No diabetes (%)	2087 (95.3)	1849 (89.3)	428 (84.6)
No hypertension (%)	1607 (73.3)	1225 (59.2)	253 (49.9)
No dyslipidaemia (%)	1912 (87.2)	1606 (77.5)	406 (80.1)
Mean SCORE (SD)	3 (5.9)	6 (10.3)	11 (12.8)
Mortality (%)	54 (2.5)	171 (8.2)	129 (25.3)



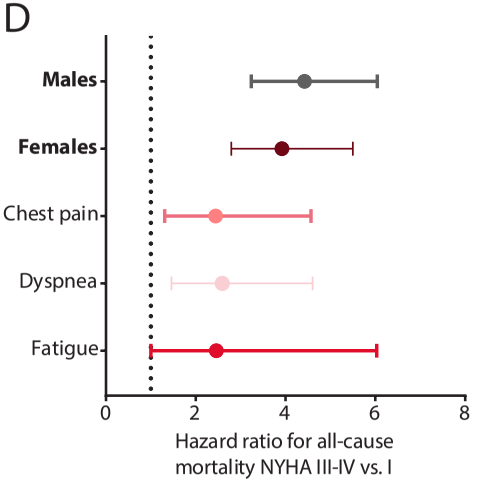


Figure 1 NYHA class in symptomatic women at the outpatient cardiology clinic A. Prevalence of complaints in women. B. Baseline table of symptomatic women stratified by NYHA-class. C. Mortality during follow up according to NYHA class. D. Hazard ratio for all-cause mortality in all men, all women and in women stratified by complaint.
BMI: body mass index, SCORE: Systematic COronary Risk Evaluation, SD: standard deviation, NYHA: New York Heart Association Classification

REFERENCES

1. Virani SS, Alonso A, Benjamin EJ, et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. *Circulation*. 2020;141(9):e139-e596. doi:10.1161/CIR.0000000000000757
2. Criteria Committee of the New York Heart Association. N*omenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels*, 9th edn. Boston: Little, Brown & Co; 1994.
3. Kajimoto K, Sato N. Sex Differences in New York Heart Association Functional Classification and Survival in Acute Heart Failure Patients With Preserved or Reduced Ejection Fraction. *Can J Cardiol*. 2020;36(1):30-36. doi:10.1016/j.cjca.2019.08.020
4. WRITING COMMITTEE MEMBERS, Yancy CW, Jessup M, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. *Circulation*. 2013;128(16):e240-e327. doi:10.1161/CIR.0b013e31829e87765
5. Ghali JK, Krause-Steinrauf HJ, Adams KF, et al. Gender Differences in Advanced Heart Failure: Insights from the BEST Study. *J Am Coll Cardiol*. 2003;42(12):2128-2134. doi:10.1016/j.jacc.2003.05.012

Chapter 5B

Sex differences in the relationship between New York Heart Association functional classification and survival in cardiovascular disease patients: A mediation analysis of exercise capacity with regular care data

Klaske R. Siegersma[¥], Niels A. Stens[¥], Floor Groepenhoff, Y. Appelman, Igor I. Tulevski, Leonard Hofstra, Hester M. den Ruijter, G. Aernout Somsen[*], N. Charlotte Onland-Moret[*]

[¥,*]These authors contributed equally

Reviews of Cardiovascular Medicine 2022; 23(8): 278

ABSTRACT

Introduction The New York Heart Association (NYHA) functional class has extensively been used for risk stratification in patients suspected of heart failure, although its prognostic value differs between sexes and disease entities. Functional exercise capacity might explain the association between the NYHA functional class and survival and can serve as an objective proxy for the subjective nature of the NYHA classification. Therefore, we assessed whether sex-differences in exercise capacity explain the association between NYHA functional class and survival in patients suspected of cardiovascular disease.

Methods For this study, we analysed electronic health record data from 7259 patients with a documented NYHA functional class and stress electrocardiogram. Follow-up for all-cause mortality was obtained through linkage with Statistics Netherlands. A mediation analysis was performed for the proportional heart rate and -workload as observed during ECG stress testing to assess to what extent these observations explain the association between NYHA functional class and survival in men and women.

Results In men, increments in NYHA functional class were related to higher all-cause mortality in a dose-response manner (NYHA II vs III/IV: hazard ratio [HR] 1.59 vs 3.64, referenced to NYHA I), whilst in women those classified as NYHA functional class II and III/IV had a similar higher mortality risk (HR 1.49 vs 1.41). The association between NYHA and survival was mostly explained by the proportional workload (men vs women: 22.9%, 95% CI 18.9%-27.3% vs 40.3%, 95% CI 28.5%-68.6%) and less so by proportional heart rate (2.5%, 95% CI 1.3%-4.3% vs 8.0%, 95% CI 4.1%-18.1% in, respectively, men and women). Post-hoc analysis showed that NYHA classification explained only a minor proportion of the association between proportional workload and all-cause mortality (15.1%, 95% CI 12.1%-18.1% vs 4.4%, 95% CI: 1.6%-7.4% in, respectively, men and women).

Conclusion This study showed a significant mediation in both sexes on the association between NYHA functional class and all-cause mortality by proportional workload, but the effect explained by NYHA classification on the association between survival and proportional workload is small. This implies that NYHA classification is not a sole representation of a patient's functional capacity, but extends to the patient's overall health status.

INTRODUCTION

The New York Heart Association (NYHA) functional classification is widely used to classify the physical limitations of patients with a variety of cardiovascular symptoms related to heart failure. Step-wise increments in the NYHA functional class were related to an increased mortality risk[1], although important sex differences were apparent. In a sex-stratified analysis of data from the BEST study, that included patients with heart failure, a NYHA class III or IV and reduced left ventricular ejection fraction, men with a NYHA class IV had a mortality risk that was almost twice as high compared to NYHA class III. In women with NYHA class IV mortality risk tripled compared to NYHA class III.[2] Registry data from patients with heart failure with reduced ejection fraction showed a similar trend with higher mortality in patients with NYHA classification IV compared to II. In these patients, NYHA class IV was a significant predictor of all-cause mortality in women, but not in men.[3] These results suggest that the NYHA classification measures disease and symptom characteristics differently in men and women.

Although originally designed for patients with heart failure[4,5], the NYHA classification is now used as a fast and easy tool for risk stratification in a large share of patients with cardiovascular symptoms visiting a physician. We previously showed that NYHA classification also has prognostic value for types of complaints other than complaints associated with heart failure.[6] Nevertheless, the evidence for risk stratification by NYHA classification in cardiovascular complaints other than heart failure remains limited.

Despite its extensive use, NYHA functional class remains a subjective method of cardiovascular disease (CVD) risk stratification[7–9], as it reflects the physician's and patient's judgment of a patient's physical condition. An aspect of the patient's physical condition is the ability to initiate and sustain exercise. This ability might explain the powerful prognostic ability of the NYHA classification.[7] Exercise capacity, i.e. the inability to achieve a maximum workload[10–13] or maximum heart rate during exercise testing[14,15], is related to an increased risk of cardiovascular disease (CVD) and all-cause mortality in men and women. Moreover, a low exercise capacity was specifically associated with CVD events in women.[16] In general, women present with a lower exercise capacity than men.[17,18] This may explain the strong prognostic value of the NYHA classification for clinical outcomes in women.

The intermediating effect of variables that represent exercise capacity on the relation between NYHA classification and all-cause mortality might provide us detailed insight in sex differences in the components of the NYHA classification. Therefore, the aim of the present study was to assess sex differences in the extent to which exercise capacity is responsible for the association between NYHA functional class and mortality risk in CVD patients.

METHODS

Study population

Electronic health record data from 2007-2018 of the Cardiology Centers of the Netherlands (CCN) were extracted. The design of the CCN database has been described before[19]. In short, the CCN network contains thirteen "one-stop shop" cardiac outpatient clinics and operates between the general practitioner and hospital cardiologist to facilitate efficient diagnostic cardiac workup. From the available 109,151 patients that were admitted to the CCN between 2007 and 2018, only patients with complete mortality data, the first documented NYHA functional class for dyspnoea, chest pain or fatigue, and stress electrocardiogram (ECG) during the same consult were selected, leaving a final study population of 7,259 patients (Figure 1).

The CCN data were made available under implied consent and transferred to the University Medical Center Utrecht under the Dutch Personal Data Protection Act. This study used data collected during the regular care process and did not subject participants to additional procedures or impose behavioural patterns on them. The Medical Research Ethics Committee of the University Medical Center Utrecht declared that research with the CCN database does not meet the Dutch Medical Research Involving Human Subjects Act (proposal number 17/359).

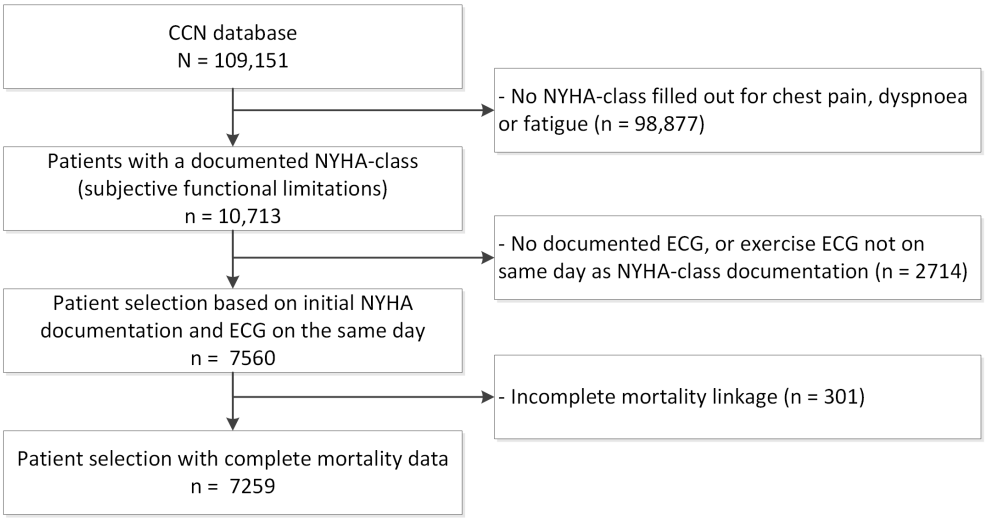


Figure 1 Flowchart of patient selection. CCN: Cardiology Centers of the Netherlands. ECG: electrocardiogram, NYHA: New York Heart Association.

Design

During a consultation, patients received a diagnostic work-up including NYHA functional class for chest pain, dyspnoea or fatigue, a detailed standardized anamnesis by a specialized nurse and cardiologist, where self-reported anthropometrics, symptoms, cardiovascular risk factors, comorbidities and medication use were registered. A NYHA classification of III or IV was converted into a combined class of NYHA class III/IV, as the number of patients documented as class IV was too small for any sensible analyses. Blood pressure measurements (Microlife WatchBP, Microlife AG, Switzerland; Medtronic BL-6 Compact, Medtronic, USA) and a 12-lead ECG (Welch Allyn Cardioperfect recorder, Welch Allyn, USA) were performed both in supine position during rest, and on a watt bike (Lode Corival Eccentric, Lode, The Netherlands) during a stress test. Predicted workload during stress was calculated based on the Jones protocol[20] and is dependent on length, age and sex. The corresponding formula is:

$$\text{Predicted workload} = (3.34*\text{Length}) - (1.43*\text{Age}) - 312 - (47*\text{Sex})$$

Sex is defined as a logical factor (i.e. women=1, men=0). Qualitative text retrieval methods were used to classify the reasons to stop the stress ECG and the conclusion of the stress ECG. The reason to stop was documented as target heart rate achieved, arrhythmia, dyspnoea, chest pain, fatigue, blood pressure and/or painful legs. The conclusion the cardiac stress ECG was documented as either normal, abnormal, inconclusive, incomplete (i.e. target heart rate not reached), myocardial infarction or arrhythmias. The variables used to define exercise capacity were calculated with the following formulas;

$$\text{Proportional heart rate} = \frac{\text{Maximum heart rate during exercise}}{\text{Predicted heart rate during exercise}}$$

$$\text{Proportional workload} = \frac{\text{Maximum workload during exercise}}{\text{Predicted workload during exercise}}$$

Follow-up for all-cause mortality was performed by linkage to Statistics Netherlands (CBS, The Hague, Netherlands; i.e. national population registry). Follow-up time was calculated as the interval between age at date of admission to the cardiology center, and age at death[21] or end of follow-up (i.e. February 2020), whichever came first.

Statistical analysis

Missing values were imputed with sex-stratified multiple imputation using the R package MICE version 3.13.0[22] with 10 imputations and 50 iterations (supplementary materials). To estimate survival function for the different NYHA classes and sexes, a time-to-event analysis using the Kaplan-Meier method and Cox proportional hazards regression was

Table 1 Baseline characteristics of included patients, stratified for sex.

	Overall	Men	Women
Total patients, n	7259	3419	3840
Age, years (SD)	57.9 (13.1)	57.2 (13.1)	58.6 (13.0)
NYHA functional class (n, %)			
I	3919 (54.0)	1913 (56.0)	2006 (52.2)
II	2908 (40.1)	1297 (37.9)	1611 (42.0)
III-IV	432 (6.0)	209 (6.1)	223 (5.8)
NYHA primary complaint (n, %)			
Chest pain	4948 (68.2)	2409 (70.5)	2539 (66.1)
Dyspnoea	1575 (21.7)	659 (19.3)	916 (23.9)
Fatigue	736 (10.1)	351 (10.3)	385 (10.0)
Positive family history, n(%)	4874 (67.1)	2149 (62.9)	2725 (71.0)
BMI, kg/m^2 (SD)	26.8 (4.9)	27.1 (4.3)	26.6 (5.4)
Smoking status (n, %)			
Never	1694 (25.2)	691 (21.7)	1003 (28.3)
Former	2364 (35.1)	1222 (38.3)	1142 (32.3)
Current	2669 (39.7)	1276 (40.0)	1393 (39.4)
Diabetes (n, %)	621 (8.6)	340 (10.0)	281 (7.3)
Hypertension (n, %)	2621 (36.1)	1230 (36.0)	1391 (36.2)
Dyslipidemia (n, %)	1333 (18.4)	671 (19.7)	662 (17.3)
Resting heart rate, beats/min (SD)	73.0 (12.4)	72.1 (12.8)	73.7 (12.0)
Arrhythmia during rest (n, %)	71 (1.2)	50 (1.9)	21 (0.7)
Medication use (n, %)			
Antihypertensive use	949 (13.1)	470 (13.7)	479 (12.5)
Cholesterol-lowering medication	511 (7.0)	284 (8.3)	227 (5.9)
Anti-diabetic medication	153 (2.1)	88 (2.6)	65 (1.7)
Anti-thrombotic medication	459 (6.3)	286 (8.4)	173 (4.5)
Anti-arrhythmic medication	22 (0.3)	<10 (<0.4)	>10 (>0.3)
Vitamin-K antagonist	66 (0.9)	34 (1.0)	32 (0.8)
Other HF medication	10 (0.1)	<10 (<0.4)	<10 (<0.3)
HeartSCORE (median [IQR])	3.4 [1.2-7.8]	3.0 [1.1-6.9]	3.7 [1.3-8.8]
Nelson-aalen estimator (median [IQR])	0.04 [0.02-0.06]	0.04 [0.02-0.06]	0.04 [0.02-0.06]
Stress ECG			
Reason to stop stress ECG (n, %)			
Target heart rate reached	1299 (17.9)	641 (18.7)	658 (17.1)
Dizziness	223 (3.1)	88 (2.6)	135 (3.5)
Fatigue	2794 (38.5)	1217 (35.6)	1577 (41.1)
Chest pain	387 (5.3)	222 (6.5)	165 (4.3)
Painful legs	2261 (31.1)	1131 (33.1)	1130 (29.4)
Arrhythmia	71 (1.0)	48 (1.4)	23 (0.6)
Dyspnoea	2341 (32.2)	948 (27.7)	1393 (36.3)
Blood pressure	258 (3.6)	160 (4.7)	98 (2.6)

	Overall	Men	Women
SBP, mmHg (SD)	199.8 (28.9)	206.1 (28.2)	194.2 (28.4)
DBP, mmHg (SD)	86.2 (20.4)	85.6 (20.6)	86.7 (20.3)
Proportional workload (SD)	0.97 (0.28)	0.86 (0.21)	1.08 (0.29)
Proportional heart rate (SD)	1.02 (0.16)	1.02 (0.16)	1.01 (0.16)
Arrhythmia during exercise (n, %)	2152 (34.9)	1139 (39.4)	1013 (31.0)
Follow-up			
All-cause mortality (n, %)	346 (4.8)	209 (6.1)	137 (3.6)
CVD mortality (n, %)	88 (1.2)	53 (1.6)	35 (0.9)
Follow-up, years (median [IQR])	5.5 [3.6-7.5]	5.5 [3.5-7.5]	5.5 [3.6-7.5]

Proportional work load and proportional heart rate are calculated proportions as described in the methods section. BMI: body mass index, CVD: cardiovascular disease, ECG: electrocardiogram, SBP: systolic blood pressure, DBP: diastolic blood pressure, IQR: interquartile range, SD: standard deviation, HeartSCORE: 10 year risk of CVD[29].

performed. Proportional hazards and linearity were verified using visual inspection of hazard function and residual plots, respectively.

To study the association between NYHA functional class and all-cause mortality, three levels of covariate adjustment were applied. The first level of adjustment was a left-truncated model that inspected the association between NYHA functional class and mortality (age-adjusted model). Secondly, a model was developed with further adjustment for known CVD risk factors and factors associated with mortality (i.e. confounder-adjusted model; supplementary materials). To identify factors associated with mortality, NYHA functional class coefficients for mortality were compared between the age-adjusted models with and without the inclusion of the variable of interest. Factors were considered confounders if they affected the NYHA functional class coefficients more than 10%. The third model, the confounder- and intermediate-adjusted model, additionally included the exercise capacity properties as intermediating variables (i.e. proportional workload and proportional heart rate).

To quantify the proportion of the association between NYHA functional class and mortality that could be explained by exercise capacity properties, we used the difference method[23,24]. Two regression coefficients of the exposure-outcome association were required: the direct effect and the total effect. The direct effect is the coefficient of the NYHA functional class in the confounder- and intermediate-adjusted model, whereas the total effect is the coefficient of the NYHA functional class in the confounder-adjusted model. The proportion of the effect explained by the intermediate (PEE) was subsequently calculated following:

$$\text{PPE} = \frac{\text{total effect - direct effect}}{\text{total effect}}$$

NYHA class II and III/IV coefficient estimates were combined via nonlinear transformation to allow the calculation of one PEE per intermediate.[25,26] Results on the different imputation sets were combined using Rubin's rules.[27] Bootstrap resampling was used to obtain 95% confidence intervals (CI) around the PEE (Supplementary Materials). Sensitivity analyses were performed per primary NYHA complaint (i.e. fatigue, dyspnoea or chest pain) and for age strata at initial consult (i.e. <65 and ≥65 years).

After evaluation of the first results, a post-hoc analysis was performed that investigated whether NYHA classification and proportional workload accounted for different aspects of risk-stratification in patients with cardiovascular complaints, which was quantified by the PEE of the NYHA functional classification for the association between the proportional workload and mortality. In this analysis, the proportional workload was set as the determinant, whereas the NYHA functional class was added to the association as the intermediating variable. The post-hoc analysis implemented proportional workload as a numerical variable multiplied by 100, converting the proportion into a percentage to ease interpretation.

In all results, NYHA classification was documented with NYHA I as the reference value. All analyses were performed in R, pooled according to Rubin's rules[28], and stratified by sex. An α-level of .05 was considered statistically significant. Data is presented as mean ± standard deviation (SD), median with interquartile range (IQR), or frequency and percentage as appropriate. All linkages and data analyses were performed within the secure environment of Statistics Netherlands, according to Dutch privacy law. Number of patients within the baseline tables and figures were occasionally too small to adequately protect privacy according to legislation regarding the use of data from Statistics Netherlands. Therefore, frequencies below 10 are presented as <10 and corresponding percentages in the data.

RESULTS

Patients had a mean age of 58 years, and 52.9% were women. Compared to men, women had overall higher NYHA functional classifications, were older (mean age 58.6 years and 57.2 years for women and men, respectively), had a lower body mass index (BMI, mean BMI 26.6 vs 27.1), and were less likely to have cardiovascular risk factors and comorbidities, e.g. were less often considered smokers (current or former smokers in women: 39.4% and 32.2% vs 40.0% and 38.3% in men), diabetic (7.3% vs 10.0% in, respectively, women and men) and dyslipidaemic (17.3% vs 19.7% in, respectively, women and men). Table 1 gives an overview of these baseline characteristics. Nonetheless, women had a higher median 10-year risk of CVD according to SCORE[29] (3.7 vs 3.0, in women and men, respectively). During both rest and stress, women were more likely to experience dyspnoea, while men were more likely to experience chest pain. Women were able to reach

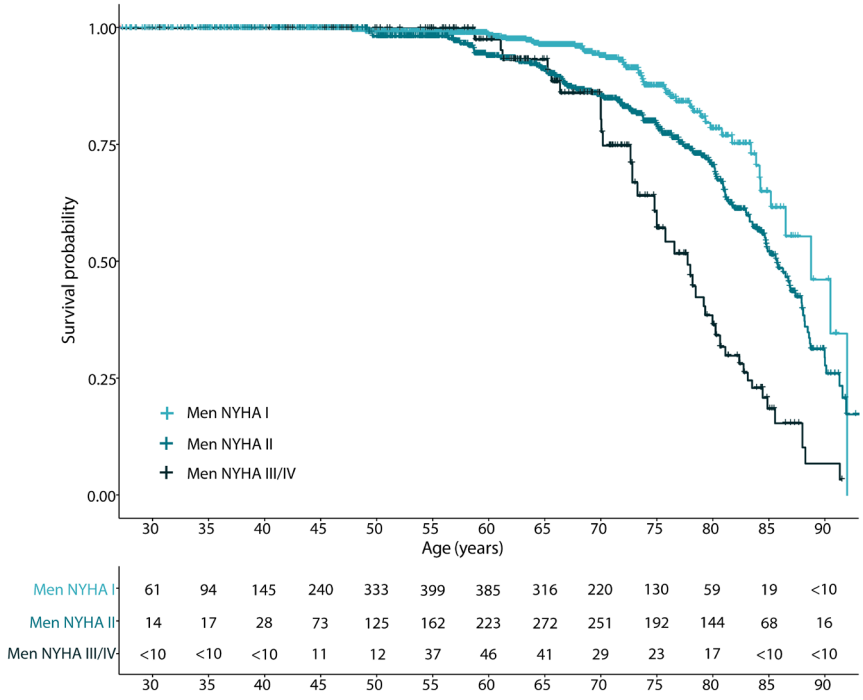


Figure 2 All-cause mortality during follow-up in men, according to NYHA functional classification.

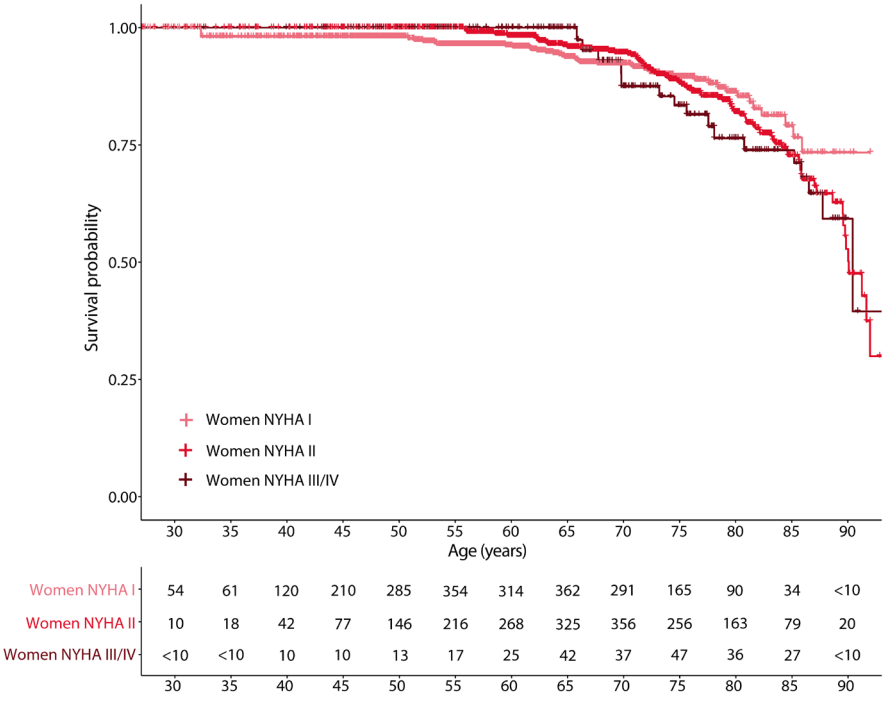


Figure 3 All-cause mortality during follow-up in women, according to NYHA classification.

Table 2 Univariate and multivariable Cox-regression analysis to evaluate the association between NYHA functional classification and all-cause mortality within men and women with cardiovascular disease.

Model	n	Event	NYHA II		NYHA III-IV	
			HR (95% CI)	p-value	HR (95% CI)	p-value
Age-adjusted model						
Men	3419	209	1.62 (1.15-2.29)	.007	3.92 (2.54-6.06)	<.001
Women	3840	137	1.50 (1.01-2.22)	.045	1.58 (0.86-2.89)	.141
Confounder model						
Men	3419	209	1.59 (1.12-2.27)	.011	3.64 (2.31-5.71)	<.001
Women	3840	137	1.49 (1.00-2.21)	.054	1.41 (0.76-2.62)	.280

HR: hazard ratio, CI: confidence interval.

a higher proportional workload despite a similar proportional heart rate compared to males (Table 1).

During a median follow-up of 5.5 years (IQR 3.6-7.5), 209 men and 137 women died. Survival analysis visualized that increments in NYHA functional class were associated with mortality in both men (Figure 2) and women (Figure 3). Univariate analysis showed that BMI (change of NYHA coefficient III/IV in respectively men and women: 9.7% and 15.8%) and conclusion of the ECG stress test (change of NYHA coefficient III/IV in respectively men and women: -7.5% and -25.9%) were confounding factors (Supplementary table 1). These variables were included in the confounding model.

The cox regression analysis confirmed that men classified as NYHA functional class II (HR 1.59, 95% CI 1.12-2.27) and NYHA functional class III/IV (HR 3.64, 95% CI 2.31-5.71) had a higher all-cause mortality risk referenced to men classified as NYHA functional class I (Table 2). Similar to men, women classified as NYHA II had a higher all-cause mortality risk than those in class I (HR 1.49, 95% CI 1.00-2.21, Table 2). Interestingly, women classified as NYHA functional class III/IV had similar mortality risks to those in class II, when compared to class I (HR 1.41, 95% CI 0.76-2.62, Table 2).

Subsequently, we extended the confounding model by adding potential intermediates for the association between NYHA functional class and all-cause mortality (Table 3). A

Table 3 Results of 1) the mediation analyses of the proportional workload and proportional heart rate on the association between NYHA classification and all-cause mortality, and 2) the post-hoc mediation analysis of NYHA classification on the association between proportional workload and mortality.

	Event rate (%)	PEE by proportional workload, % (95% CI)	PEE by proportional heart rate, % (95% CI)	PEE by NYHA classification, % (95% CI)
Men	6.1	22.9 (18.9-27.3)	2.5 (1.3-4.3)	15.1 (12.1-18.1)
Women	3.6	40.3 (28.5-68.6)	8.0 (4.1-18.1)	4.4 (1.6-7.4)

PEE: proportion effect explained, CI: confidence interval.

Table 4 Univariate and multivariable Cox-regression analysis to evaluate the association between the proportional workload and mortality within men and women with cardiovascular disease.

Model	n	Event	Proportional workload	
			HR (95% CI)	p-value
Age-adjusted model				
Men	3419	209	0.973 (0.966-0.981)	<.001
Women	3840	137	0.985 (0.979-0.991)	<.001
Confounder model				
Men	3419	209	0.974 (0.966-0.982)	<.001
Women	3840	137	0.988 (0.982-0.994)	<.001

HR: hazard ratio, CI: confidence interval.

statistically significant, but small proportion of this association between NYHA and mortality was explained by the proportional heart rate, being more profound in women than in men (men vs. women: 2.5%, 95% CI 1.3%-4.3% vs 8.0%, 95% CI 4.1%-18.1%). A stronger pattern was observed for the proportional workload (men vs women: 22.9%, 95% CI 18.9%-27.3% vs 40.3%, 95% CI 28.5%-68.6%).

The post-hoc analysis showed that lowering in proportional workload was associated with a higher mortality risk (Table 4), in men (HR per % lowering in proportional load of the age-adjusted model: 0.973, 95% CI 0.966-0.981, HR confounder-adjusted model: 0.974, 95% CI 0.966-0.981) and to a lesser extent in women (HR age-adjusted model: 0.985, 95% CI 0.979-0.991, HR confounder-adjusted model: 0.988, 95% CI 0.982-0.994). The mediation analysis showed that only a minor proportion of the association between proportional workload and mortality was explained by NYHA functional class in both men and women (15.1%, 95% CI 12.1%-18.1 vs 4.4%, 95% CI 1.6%-7.4%, respectively). Supplementary table 1 shows the results of the univariate analysis of confounders.

Sensitivity analyses performed to elucidate the influence of age at initial consult and the primary complaint led to similar conclusions. Increments in NYHA functional class were related to all-cause mortality in both men aged <65 and ≥65 years, whilst this trend was absent in women in both age-groups (Supplementary table 2). When stratified by primary complaint, step-wise increases in NYHA functional class were significantly associated with all-cause mortality in men, but not in women (Supplementary table 3).

DISCUSSION

The aim of the present study was to assess the extent to which exercise capacity properties in men and women separately are responsible for the association between NYHA functional class and all-cause mortality in CVD patients. We first showed that increments in NYHA functional class were related to all-cause mortality risk in both men and women that underwent stress testing, although this seemed to be stronger in men than in women. Second, the proportional workload explained a significant proportion of the as-

sociation between NYHA functional class and all-cause mortality in men and women, although the majority of this association remained unexplained. Third, the post-hoc analysis showed a lower PEE of NYHA classification in the association between proportional workload and survival compared to the PEE of proportional workload in the association between NYHA classification and survival. Taken together, these results suggest that the NYHA functional class and exercise test provide distinct information within the clinical risk assessment of men and women.

For the current study, we used a unique and large population of patients presenting with a wide variety of symptoms that were admitted to the CCN; an outpatient cardiology clinic which operates between the general practitioner and the hospital. This set-up leads to a population that closely resembles the population with cardiovascular symptoms at the general practitioner's office. For example, within the current study population, ~53% of the admitted patients were women, providing a solid basis for investigating sex differences within this population. In addition, all centers of the CCN network follow a standardized diagnostic workflow during each consultation, resulting in a high-quality and structured data collection.

The presented study has several limiting factors. First, there are limitations in the selected study population. To enable mediation analysis, only individuals with a documented ECG stress test were selected. This resulted in the exclusion of mainly older women, who suffered from dyspnoea and were classified as NYHA functional class III/IV in whom no ECG stress test was performed. We did not replicate the high prognostic value of the NYHA functional class, especially in NYHA class III/IV, in women that we previously observed[6] whilst sampling from the same population. This suggests that some extent of collider bias was introduced in the presented study by conditioning on the presence of the ECG stress test. This specific selection led to a healthier selected female population with CVD, which distorted the survival estimates in women. In addition, this specific selection prevents accurately estimating the underlying NYHA distribution within this population. Another disadvantage of the selected population is that it also includes patients that have not reached their target heart rate during ECG stress testing. Although these patients have an invalid stress test, exclusion of this population might lead to even more bias as only the very healthy patients were included. Second, although medication use did not differ between men and women, data regarding subsequent treatment was not completely captured. We can therefore not exclude potential sex differences during follow-up, which may have affected all-cause mortality rates in men and women and its relation with NYHA functional class. Third, although all centers of the CCN followed a standardized diagnostic workflow, some cardiologists deviated from the stress ECG protocol due to instability of the patient, which may have influenced estimates of the proportional workload and heart rate and their subsequent PEE. Fourth, only the first-doc-

umented NYHA functional class of the patient was selected, which was generally during their initial consult (documented NYHA classification consult, median 1, IQR 1-1). However, NYHA functional classification during follow-up may fluctuate in response to disease progression or treatment, which may have affected our hazard ratios in either direction. Finally, the retrospective and observational design, despite adopting a multivariate analysis, may also have affected our survival estimates[30].

The NYHA functional class is extensively being used in clinics for a wide variety of applications, including clinical trial inclusion criteria, disease management and prognosis[31,32]. Previous studies highlighted that increments in NYHA functional class were related to all-cause mortality in both men and women with heart failure with preserved ejection fraction[3], but only in women with reduced ejection fraction[2,3]. We previously highlighted that increments in NYHA functional class were associated with all-cause mortality in both men and women with CVD.[6] In contrast, the present study, that sampled from the same population[6], showed that stepwise increases in NYHA functional class were related to all-cause mortality risk in men, whilst in women the mortality risk was similar among those classified as NYHA functional class II and III/IV. The introduction of collider bias may therefore have affected the survival estimates of women, although it remains unclear whether this also influenced our PEE estimates obtained in the mediation analysis. We hypothesize that, if these older, excluded women classified as NYHA class III/IV presented with complete stress ECG data, this might have resulted in an overall lower proportional workload in women. Subsequently, a larger proportion of the association between NYHA functional class and all-cause mortality may be explained by the proportional workload in women. Future studies are needed to confirm these hypotheses.

Prior studies have tried to objectify the subjective nature of the NYHA functional class by focusing on exercise[7,33–35], and showed that increments in NYHA functional class inversely correlate with objective measures of exercise capacity[33–35]. Within the present study, the proportional workload explained a significant proportion of the association between NYHA functional class and all-cause mortality in both men and women (22.9% vs 40.3%, respectively), although a large part of this association remained unexplained by variables that represent exercise capacity. In addition, only a minor proportion of the association between the proportional workload and all-cause mortality was explained by NYHA functional class (men vs women: 15.1% vs 4.4%). These results together suggest that the NYHA functional class and ECG stress testing focus on distinct elements within the CVD risk assessment. This has already been hinted at, as previous evidence demonstrated that NYHA functional class poorly differentiated across the spectrum of functional impairment[36–38]. It may therefore be advised to use an ECG stress test as an extension of the NYHA functional class for clinical risk assessment, rather than as a direct replacement. Furthermore, large differences in PEE estimates were observed in men and women sig-

nifying that the NYHA functional class does not focus on the same disease and symptom characteristics of the risk assessment in men and women. The origin of this discrepancy remains to be elucidated, but we can address the following points. First, differences in presentation of symptoms may prevent uniform classification of NYHA functional class among sexes, as women more often report atypical symptoms[39–42] and concurrent depressive symptoms[43] than men. In addition, sex-discordance between the patient and treating physician may influence symptom perception[44] and risk stratification for clinical outcomes[45–48]. Unfortunately, we were unable to assess sex-discordances within the present study, which therefore cannot be ruled out. Finally, it seems that women suffer more from functional impairments than men, which is suggested by the larger PEE estimate of the proportional workload in the association between NYHA functional class and all-cause mortality in women. Sex differences in CVD-induced adaptations in cardiac structure[1,48–50] may be the cornerstone of these more pronounced functional impairments in women. The differential domains of the NYHA functional class in men and women, paired with its inherent subjective nature, question its reliability within the clinical risk assessment. Nonetheless, the NYHA functional class remains an important prognostic tool for clinical outcomes in both men and women, and cannot directly be replaced by objective variables that represent exercise capacity. This warrants future research to further elaborate on the different domains of the NYHA functional class in men and women.

CONCLUSION

This study showed a significant mediation in both sexes on the association between NYHA functional class and all-cause mortality by proportional workload. The effect explained by NYHA classification on the association between survival and proportional workload is small. This implies that the NYHA classification is not a sole representation of the patient's functional capacity, but extends to the patient's overall health status. Although the subjective NYHA functional class tends to focus on different domains among sexes, it remains an easy-to-apply and important prognostic tool of CVD risk stratification in both men and women.

REFERENCES

1. Frazier CG, Alexander KP, Newby LK, et al. Associations of Gender and Etiology With Outcomes in Heart Failure With Systolic Dysfunction. A Pooled Analysis of 5 Randomized Control Trials. *J Am Coll Cardiol*. 2007;49(13):1450-1458. doi:10.1016/j.jacc.2006.11.041
2. Ghali JK, Krause-Steinrauf HJ, Adams KF, et al. Gender Differences in Advanced Heart Failure: Insights from the BEST Study. *J Am Coll Cardiol*. 2003;42(12):2128-2134. doi:10.1016/j.jacc.2003.05.012
3. Kajimoto K, Sato N. Sex Differences in New York Heart Association Functional Classification and Survival in Acute Heart Failure Patients With Preserved or Reduced Ejection Fraction. *Can J Cardiol*. 2020;36(1):30-36. doi:10.1016/j.cjca.2019.08.020
4. Bennett JA, Riegel B, Bittner V, Nichols J. Validity and reliability of the NYHA classes for measuring research outcomes in patients with cardiac disease. *Heart Lung*. 2002;31(4):262-270. doi:10.1067/mhl.2002.124554
5. White PD, Myers MM. The classification of cardiac diagnosis. *JAMA*. 1921;77(18):1414-1415.
6. Siegersma KR, Groepenhoff F, Onland-Moret NC, et al. New York Heart Association class is strongly associated with mortality beyond heart failure in symptomatic women. *Eur Hear J - Qual Care Clin Outcomes*. 2021;7(2):214-215. doi:10.1093/ehjqcco/qcaa091
7. Raphael C, Briscoe C, Davies J, et al. Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure. *Heart*. 2007;93(4):476-482. doi:10.1136/hrt.2006.089656
8. Goldman L, Hashimoto B, Cook EF, Loscalzo A. Comparative reproducibility and validity of systems for assessing cardiovascular functional class: Advantages of a new specific activity scale. *Circulation*. 1981;64(6):1227-1234. doi:10.1161/01.CIR.64.6.1227
9. Goode KM, Nabb S, Cleland JGF, Clark AL. A Comparison of Patient and Physician-Rated New York Heart Association Class in a Community-Based Heart Failure Clinic. *J Card Fail*. 2008;14(5):379-387. doi:10.1016/j.cardfail.2008.01.014
10. Laukkanen JA, Rauramaa R, Kurl S. Exercise workload, coronary risk evaluation and the risk of cardiovascular and all-cause death in middle-aged men. *Eur J Cardiovasc Prev Rehabil*. 2008;15:285-292. doi:10.1111/j.1365-2796.2008.02006.x
11. Smith LV, Myc L, Watson D, Beller GA, Bourque JM. A high exercise workload of ≥ 10 METS predicts a low risk of significant ischemia and cardiac events in older adults. *J Nucl Cardiol*. 2020;27(5):1486-1496. doi:10.1007/s12350-018-1376-7
12. Brawner CA, Abdul-Nour K, Lewis B, et al. Relationship between Exercise Workload during Cardiac Rehabilitation and Outcomes in Patients with Coronary Heart Disease. *Am J Cardiol*. 2016;117(8):1236-1241. doi:10.1016/j.amjcard.2016.01.018
13. Gulati M, Black HR, Shaw LJ, et al. The Prognostic Value of a Nomogram for Exercise Capacity in Women. *N Engl J Med*. 2005;353:468-475.
14. Kiviniemi AM, Tulppo MP, Hautala AJ, et al. Long-term outcome of patients with chronotropic incompetence after an acute myocardial infarction. *Ann Med*. 2011;43(1):33-39. doi:10.3109/07853890.2010.521764
15. Savonen KP, Lakka TA, Laukkanen JA, et al. Heart rate response during exercise test and cardiovascular mortality in middle-aged men. *Eur Heart J*. 2006;27(5):582-588. doi:10.1093/eurheartj/ehi708
16. Korpelainen R, Lämsä J, Kaikkonen KM, et al. Exercise capacity and mortality – a follow-up study of 3033 subjects referred to clinical exercise testing. *Ann Med*. 2016;48(5):359-366. doi:10.1080/07853890.2016.1178856

17. Ghiselli L, Marchi A, Fumagalli C, et al. Sex-related differences in exercise performance and outcome of patients with hypertrophic cardiomyopathy. *Eur J Prev Cardiol*. 2020;27(17):1821-1831. doi:10.1177/2047487319886961
18. Harms CA, Rosenkranz S. Sex differences in pulmonary function during exercise. *Med Sci Sports Exerc*. 2008;40(4):664-668. doi:10.1249/MSS.0b013e3181621325
19. Bots SH, Siegersma KR, Onland-Moret NC, et al. Routine clinical care data from thirteen cardiac outpatient clinics: design of the Cardiology Centers of the Netherlands (CCN) database. *BMC Cardiovasc Disord.* 2021;21(1):1-9. doi:10.1186/s12872-021-02020-7
20. Jones NL. *Clinical Exercise Testing*. Saunders; 1997. https://books.google.nl/books?id=hxRsAAAAMAAJ
21. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am J Epidemiol*. 1997;145(1):72-80. doi:10.1093/oxfordjournals.aje.a009309
22. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1-67. doi:10.18637/jss.v045.i03
23. Judd CM, Kenny DA. Process analysis: Estimating Mediation in Treatment Evaluations. *Eval Rev*. 1981;5(5):602-619. doi:10.1177/0193841X8100500502
24. Baron RM, Kenny DA. The Moderator-Mediator Variable Distinction in Social Psychological Research. Conceptual, Strategic, and Statistical Considerations. *J Pers Soc Psychol*. 1986;51(6):1173-1182. doi:10.1037/0022-3514.51.6.1173
25. Hayes AF, Preacher KJ. Statistical mediation analysis with a multicategorical independent variable. *Br J Math Stat Psychol*. 2014;67(3):451-470. doi:10.1111/bmsp.12028
26. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods.* 2008;40(3):879-891. doi:10.3758/BRM.40.3.879
27. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Vol 81. John Wiley \& Sons; 2004.
28. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Stat Med*. 2013;32(26):4499-4514. doi:10.1002/sim.5844
29. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J*. 2003;24(11):987-1003. doi:10.1016/S0195-668X(03)00114-3
30. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *Am J Epidemiol*. 2007;166(6):646-655. doi:10.1093/aje/kwm165
31. Criteria Committee of the New York Heart Association. *Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels*, 9th edn. Boston: Little, Brown & Co; 1994.
32. Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J*. 2016;37(27):2129-2200m. doi:10.1093/eurheartj/ehw128
33. Yap J, Lim FY, Gao F, Teo LL, Lam CSP, Yeo KK. Correlation of the New York Heart Association Classification and the 6-Minute Walk Distance: A Systematic Review. *Clin Cardiol.* 2015;38(10):621-628. doi:10.1002/clc.22468
34. Das BB, Young ML, Niu J, Mendoza LE, Chan KC, Roth T. Relation Between New York Heart Association Functional Class and Objective Measures of Cardiopulmonary Exercise in Adults With Congenital Heart Disease. *Am J Cardiol*. 2019;123(11):1868-1873. doi:10.1016/j.amjcard.2019.02.053
35. Russell SD, Saval MA, Robbins JL, et al.

New York Heart Association functional class predicts exercise parameters in the current era. *Am Heart J*. 2009;158(4 SUPPL.):S24-S30. doi:10.1016/j.ahj.2009.07.017
36. Caraballo C, Desai NR, Mulder H, et al. Clinical Implications of the New York Heart Association Classification. *J Am Heart Assoc*. 2019;8(23):1-6. doi:10.1161/JAHA.119.014240
37. Rostagno C, Galanti G, Comeglio M, Boddi V, Olivo G, Gastone Neri Serneri G. Comparison of different methods of functional evaluation in patients with chronic heart failure. *Eur J Heart Fail*. 2000;2(3):273-280. doi:10.1016/S1388-9842(00)00091-X
38. Smith RF, Johnson G, Ziesche S, Bhat G, Blankenship K, Cohn JN. Functional capacity in heart failure. Comparison of methods for assessment and their relation to other indexes of heart failure. The V-HeFT VA Cooperative Studies Group. *Circulation*. 1993;87(6 Suppl):VI88-93.
39. Canto JG, Goldberg RJ, Hand MM, et al. Symptom presentation of women with acute coronary syndromes. *Arch Intern Med*. 2007;167(22):2405-2413. doi:10.1097/IEB.0b013e31816c4230
40. Canto JG, Rogers WJ, Goldberg RJ, et al. Association of age and sex with myocardial infarction symptom presentation and in-hospital mortality. *JAMA*. 2012;307(8):813-822. doi:10.1001/jama.2012.199
41. Keteepe-Arachi T, Sharma S. Management of Refractory Angina Pectoris Ischaemic Heart Disease. *Eur Cardiol Rev*. 2017;1(12):10-13. doi:10.15420/ecr.2016
42. Shin JY, Martin R, Suls J. Meta-analytic evaluation of gender differences and symptom measurement strategies in acute coronary syndromes. *Heart Lung*. 2010;39(4):283-295. doi:10.1016/j.hrtlng.2009.10.010
43. Bucciarelli V, Caterino AL, Bianco F, et al. Depression and cardiovascular disease: The deep blue sea of women's heart. *Trends Cardiovasc Med*. 2020;30(3):170-176. doi:10.1016/j.tcm.2019.05.001
44. Okunrintemi V, Valero-Elizondo J, Patrick B, et al. Gender differences in patient-reported outcomes among adults with atherosclerotic cardiovascular disease. *J Am Heart Assoc*. 2018;7(24). doi:10.1161/JAHA.118.010498
45. Gross R, McNeill R, Davis P, Lay-Yee R, Jatrana S, Crampton P. The association of gender concordance and primary care physicians' perceptions of their patients. *Women Heal*. 2008;48(2):123-144. doi:10.1080/03630240802313464
46. Greenwood BN, Carnahan S, Huang L. Patient-physician gender concordance and increased mortality among female heart attack patients. *Proc Natl Acad Sci U S A*. 2018;115(34):8569-8574. doi:10.1073/pnas.1800097115
47. Tsugawa Y, Jena AB, Figueroa JF, Orav EJ, Blumenthal DM, Jha AK. Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians. *JAMA Intern Med*. 2017;177(2):206-213. doi:10.1001/jamainternmed.2016.7875
48. Cioffi G, Stefenelli C, Tarantini L, Opasich C. Prevalence, predictors, and prognostic implications of improvement in left ventricular systolic function and clinical status in patients >70 years of age with recently diagnosed systolic heart failure. *Am J Cardiol*. 2003;92(2):166-172. doi:10.1016/S0002-9149(03)00532-0
49. Cuocolo A, Sax FL, Brush JE, Maron BJ, Bacharach SL, Bonow RO. Left ventricular hypertrophy and impaired diastolic filling in essential hypertension. Diastolic mechanisms for systolic dysfunction during exercise. *Circulation*. 1990;81(3):978-986. doi:10.1161/01.CIR.81.3.978
50. Gori M, Lam CSP, Gupta DK, et al. Sex-specific cardiovascular structure and function in heart failure with preserved ejection fraction. *Eur J Heart Fail*. 2014;16(5):535-542. doi:10.1002/ejhf.67

SUPPLEMENTARY MATERIALS

Multiple imputation

Missing values were imputed using the R package MICE 3.13.0[1] with 10 imputations and 50 iterations. Multiple imputation was performed for each sex individually to account for effect modification between NYHA functional class and sex. Predictive mean matching was used for continuous variables (i.e. body mass index, systolic and diastolic blood pressure during stress, resting heart rate, proportional heart rate and -workload), logistic regression for binary variables (i.e. presence of diabetes mellitus, dyslipidaemia, positive family history and heart rhythm abnormalities during rest and stress), multinomial logit models for unordered categorical variables with more than 2 levels (i.e. ECG diagnosis: being either normal, abnormal, or inconclusive), and ordered logit models for ordered categorical variables with more than 2 levels (i.e. smoking status). In addition, we included age at initial consult and at event[2], as age represented the underlying time-scale for the Cox proportional hazard regression models; a Nelson-Aalen[3] estimator as well as all-cause mortality as our outcome variable. All predictors were checked for correlation (-0.7 or 0.7), but no significant correlation was observed.

Univariate analysis

The age-adjusted model was extended with one of the covariates to quantify the relation of the corresponding covariate with all-cause mortality, and was defined as the percentual change of the NYHA coefficients. Factors were considered influential if they changed the NYHA coefficients more than 10% when compared to the age-adjusted model. Although chest pain and dyspnoea during rest and during exercise were considered influential (Supplementary table 1), these covariates were not incorporated in the confounding model as the NYHA functional classification is a reflection of these cardiac symptoms. Based on univariate analysis, BMI and conclusion of the stress ECG were selected as relevant confounders (Supplementary table 1). These covariates were extended with known CVD risk factors (i.e. diabetes and family history), together constituting the confounding model.

Bootstrapping

Bootstrapping (1000 bootstrap samples) was performed to compute 95% confidence intervals around the PEE. Bootstrapping was performed per sex and for every imputation set individually. Regression coefficients with and without the intermediate of interest were determined in each imputation set. For each individual bootstrap, obtained total and direct effects of the ten imputation sets were pooled per sex according to Rubin's Rule. Finally, 1000 PEE's were obtained of which the 0.025^{th} and 0.975^{th} percentile were taken as the 95% confidence interval around the PEE.

REFERENCES

1. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw.* 2011;45(3):1-67. doi:10.18637/jss.v045.i03
2. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am J Epidemiol*. 1997;145(1):72-80. doi:10.1093/oxfordjournals.aje.a009309
3. White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med.* 2009;28:1982-1998. doi:10.1002/sim.3618

Supplementary Tables

Supplementary table 1 Univariate associations with all-cause mortality.

	NYHA II		NYHA III-IV		Proportional workload	
	Men	Women	Men	Women	Men	Women
Age	REF	REF	REF	REF	REF	REF
BMI	4.66	2.85	9.70	15.83	0.54	0.19
Smoking status	-3.43	-0.19	-1.26	0.97	-0.70	2.15
Positive family history	0.79	0.38	1.44	2.61	0.05	-0.03
Diabetes	-0.39	-0.42	-4.96	-6.31	0.40	-2.93
Dyslipidemia	-1.79	-0.62	2.44	-1.08	-0.17	0.29
Hypertension	0.24	4.35	0.27	2.88	0.00	0.73
Complaints during rest	-12.35	-6.66	-37.40	-54.00	-5.30	-7.66
Chest pain	-16.28	-10.94	-34.00	-39.76	-6.65	-3.88
Dyspnoea	3.61	2.84	-3.69	-8.40	3.49	-1.33
Fatigue	0.32	0.90	2.83	3.48	-0.10	-0.07
Number of complaints						
Medication use						
Antihypertensive use	-2.85	-1.01	0.47	0.27	-0.37	0.09
Cholesterol-lowering medication	2.96	0.94	3.75	1.47	0.72	1.22
Anti-diabetic medication	0.53	-0.18	-1.48	-1.93	1.17	-0.53
Anti-thrombotic medication	-1.85	-1.03	-0.97	-0.06	0.46	-0.70
Anti-arrhythmic medication	-1.52	-0.16	-0.53	0.61	0.98	-0.11
Vitamin-K antagonist	-3.68	0.57	0.73	-0.51	-0.30	0.06
Other HF medication	-1.71	-0.26	0.76	-0.01	-1.80	0.27
Resting heart rate	-0.65	-0.95	-0.99	-1.57	-1.34	-0.03
Arrhythmia during rest	-4.18	-1.54	-4.97	-7.75	-1.66	-2.23
Stress ECG						
SBP	-3.85	-7.88	0.15	-1.61	-10.36	2.37
DBP	-0.78	-0.37	4.68	-1.31	0.11	0.91
Arrhythmia during stress	0.15	-0.05	-2.15	0.41	-0.15	2.32
ECG diagnosis	-6.47	-4.32	-7.53	-25.89	-3.38	-17.80
Reason to stop						
Heart rate	-2.59	-2.47	-5.36	-8.84	-0.04	-4.12
Dizzyness	-1.23	-0.80	-0.03	-1.31	-0.69	-0.19
Fatigue	-2.68	-0.24	-0.41	0.50	1.05	0.16
Chest pain	3.37	3.24	-1.26	0.28	0.11	-0.43
Painful legs	-2.37	-0.62	1.75	2.18	0.22	-0.22
Arrhythmia	-0.12	0.33	0.14	0.49	0.29	0.11
Dyspnoea	-3.88	-4.90	-9.81	-22.68	-2.26	0.20
Blood pressure	-1.79	-1.43	0.08	-0.98	-0.54	-0.31

BMI: body mass index, ECG: electrocardiogram, HF: heart failure, HR: heart rate, REF: reference.

Supplementary table 2 Multivariate analysis for men and women with cardiovascular disease, stratified for age at initial consult.

Model	n	Event	NYHA II		NYHA III-IV	
			HR (95% CI)	p-value	HR (95% CI)	p-value
Age-adjusted model						
Men, <65 years	2391	47	3.00 (1.56-5.76)	.002	4.23 (1.62-11.04)	.005
Women, <65 years	2509	36	0.82 (0.41-1.64)	.571	1.14 (0.26-4.87)	.865
Men, ≥65 years	1028	162	1.29 (0.87-1.91)	.213	3.51 (2.15-5.73)	<.001
Women, ≥65 years	1331	101	2.11 (1.25-3.57)	.006	2.09 (1.02-4.28)	.046
Confounder model						
Men, <65 years	2391	47	3.30 (1.67-6.51)	.001	4.60 (1.69-12.51)	.005
Women, <65 years	2509	36	0.88 (0.43-1.78)	.722	1.22 (0.28-5.40)	.795
Men, ≥65 years	1028	162	1.26 (0.84-1.89)	.261	3.39 (2.05-5.60)	<.001
Women, ≥65 years	1331	101	2.14 (1.25-3.64)	.006	2.07 (0.99-4.34)	.057

HR: hazard ratio, CI: confidence interval, NYHA: New York Heart Association functional classification

Supplementary table 3 Multivariate analysis for men and women with cardiovascular disease, stratified for primary NYHA complaint.

Model	n	Event	NYHA II		NYHA III-IV	
			HR (95% CI)	p-value	HR (95% CI)	p-value
Age-adjusted model						
Men, chest pain	2409	114	1.47 (0.96-2.26)	.078	2.88 (1.51-5.50)	.002
Women, chest pain	2539	59	1.36 (0.78-2.38)	.285	2.24 (0.84-5.95)	.112
Men, dyspnoea	659	78	1.50 (0.75-2.99)	.255	4.08 (1.88-8.85)	.001
Women, dyspnoea	916	60	0.86 (0.42-1.74)	.677	0.55 (0.20-1.56)	.267
Men, fatigue	351	17	1.43 (0.29-7.11)	.671	4.22 (0.73-24.47)	.132
Women, fatigue	385	18	1.14 (0.34-3.81)	.830	3.27 (0.84-12.81)	.110
Confounder model						
Men, chest pain	2409	114	1.49 (0.95-2.32)	.084	2.94 (1.52-5.69)	.002
Women, chest pain	2539	59	1.38 (0.78-2.45)	.272	2.20 (0.79-6.13)	.138
Men, dyspnoea	659	78	1.45 (0.72-2.93)	.303	3.72 (1.67-8.31)	.002
Women, dyspnoea	916	60	0.82 (0.40-1.68)	.594	0.49 (0.17-1.40)	.189
Men, fatigue	351	17	1.36 (0.26-7.01)	.724	5.86 (0.94-36.49)	.093
Women, fatigue	385	18	0.97 (0.27-3.51)	.969	2.74 (0.63-11.95)	.213

HR: hazard ratio, CI: confidence interval, NYHA: New York Heart Association functional classification

Chapter 6

Development of a pipeline for adverse drug reaction identification in clinical notes: word embedding models and string matching

Klaske R. Siegersma[¥], Maxime Evers, Sophie H. Bots, Floor Groepenhoff, Yolande Appelman, Leonard Hofstra, Igor I. Tulevski, G. Aernout Somsen, Hester M. den Ruijter, Marco Spruit[*], N. Charlotte Onland-Moret[*]

[*] These authors contributed equally

JMIR Medical Informatics 2022; 10(1): 1-13

ABSTRACT

Background Knowledge about adverse drug reactions (ADRs) in the population is limited due to underreporting, which hampers surveillance and assessment of drug safety. Therefore, gathering accurate information about incidence of ADRs is of great relevance, which can be retrieved from clinical notes. However, manual labelling of these notes is time-consuming and automatization can improve use of free text clinical notes for identification of ADRs. Furthermore, tools for language processing in languages other than English are not widely available.

Objective The aim of this study is to design and evaluate a method for automatic extraction of medication and Adverse Drug Reaction Identification in Clinical Notes (ADRIN).

Methods Dutch free text clinical notes (n=277.398) and medication registrations (n=499.435) were used from the Cardiology Centers of the Netherlands database. All clinical notes were used to develop word embedding models. Vector representations of word embedding models and a string matching with a medical dictionary (MedDRA) were used for identification of ADRs and medication in a test set of clinical notes that was manually labelled. Several settings, including search area and punctuation, could be adjusted in the prototype to evaluate the optimal version of the prototype.

Results The ADRIN method was evaluated using a test set 988 clinical notes, written on the stop date of a drug. Multiple versions of the prototype were evaluated for a variety of tasks. Binary classification of ADR presence achieved the highest accuracy of 0.84. Reduced search area and inclusion of punctuation improved performance, while incorporation of MedDRA did not improve performance of the pipeline.

Conclusions The ADRIN method and prototype are effective in recognizing ADRs in Dutch clinical notes from cardiac diagnostic screening centers. Surprisingly, incorporation of MedDRA did not result in improved identification on top of word embedding models. The implementation of the ADRIN tool may help to increase the identification of ADRs, resulting in better care and saving substantial healthcare costs.

INTRODUCTION

Literature shows that adverse drug events (ADEs) and more specifically adverse drug reactions (ADRs) are structurally underreported.[1] Clinical trials may underreport or miss ADRs for various reasons, such as a follow-up that is usually too short to catch long-term effects.[2] In addition, the study population may be healthier or otherwise different from the target population in regular care.[3] As a result, the ADR risk of clinically relevant subgroups such as women and the elderly remain unknown[4], which places a societal and economic burden on our healthcare system. The prevalence of hospital admissions associated with ADRs is reported to be as high as 5.3% and is estimated to be twice as high for the older adult population.[5] In the United States alone ADRs are estimated to generate 30 billion dollars of unnecessary costs.[6] Yet, efforts have been made to structurally collect information on ADRs, both on a national (e.g. Lareb in the Netherlands) and international (EudraVigilance[7]) level. However, these pharmacovigilance databases do not include relevant patient characteristics and information about prescription rates.

Regular care data extracted from electronic health records (EHR) can help in post-marketing surveillance of medication. ADRs are usually not reported in the EHR in a structured way, but the clinical notes made during consultations between patients and their physician may hold relevant information when patients experienced an ADR. However, these notes are often stored as free text and thus cannot be easily analyzed.[8] Methods that extract ADRs from these free text fields are needed to access the full potential of these data.

Natural language processing (NLP) techniques can aid in differentiation of relevant features from idle free text and prepare free text for research purposes.[9,10] One of the widespread topics in NLP is the use of word embeddings; a vector representation of a text, often established through evaluation of the word's context. The use of word embeddings for evaluation of clinical free text for research purposes is increasing.[11] Research has shown that training word embedding models on a domain-specific dataset generates better results than training on a general dataset.[12,13] As a result, applications of word embedding models are studied in a wide range of topics within the healthcare domain; e.g. evaluation of radiology reports[14], identification of ICD-10 codes[15], identification of adverse drug events in English EHRs[16] and can potentially be a solution to extract ADRs from Dutch clinical notes.

The objective of this research was to design a method for the identification of ADRs in clinical notes from a regular care database (ADRIN; Adverse Drug Reactions Identification in clinical Notes) using unlabeled data and word embeddings. Although demonstrations in this study were done with Dutch clinical notes from the cardiovascular domain, the method was developed in a way that not only enables generalization to other languages, but also to other research questions to mine text in clinical notes.

METHODS

The ADRIN method is based upon implementation of a medical taxonomy to enhance standardized terminology (Medical Dictionary for Regulatory Activities, MedDRA[17]) and word embeddings, trained on a large database of medical free text. Additionally, a prototype was developed and evaluated on labelled Dutch clinical notes to determine the performance of this method. Figure 1 demonstrates the general workflow of the ADRIN method.

This study focussed on the identification of ADRs and the corresponding medication. We assume that patients were compliant to their medication regimen. We defined an ADR as any unwanted event that led to the discontinuation of the prescribed medication. In the following description, clinical notes are defined as the free text written down in the EHR by the physician after a patient's consult.

Dataset

The Cardiology Centres of the Netherlands (CCN) is a large, regular care database from thirteen diagnostic cardiac screening centres. In short, this database consists of 109,151 patients that visited one of the outpatient cardiac screening centres between 2007 and 2018 and includes patient characteristics and information about diagnostic tests.[18]

In total, there were 277,398 clinical notes in the database and 499,435 medication prescriptions. Clinical notes were de-identified with DEDUCE-software.[19] Medication prescriptions contain information about the prescribed medication, start date and, if the medication was discontinued at some point, end date and reason for discontinuation in free text.

Figure 2 describes the selection of discontinued medication entries from the database. The selected prescriptions were merged with the clinical notes. This resulted in 91,273 discontinued medication entries for which a clinical note was available on the end date of the medication. In cases where multiple prescriptions from the same patient were stopped on the same day (n=19,992), the same clinical note was used for all prescriptions. A reason for discontinuation was reported in 36,508 (39.9%) medication prescriptions. We randomly selected 1000 medication entries and corresponding clinical notes as a test set from these medication entries. However, in 12 cases the clinical note was empty, resulting in a test set of 988 clinical notes.

The validation set was obtained from discontinued medication entries and consisted of all medication stops with an ADR reported as a reason for discontinuation and a random selection of 1,600 of medication stops that were not related to an ADR. The latter selection was made as it was, because we expected that clinical notes corresponding to these medication stops, might also contain information on possible ADRs. This selection made it thus more likely that medication and ADRs would be identified, when compared

to a random selection of all clinical notes (Figure 2). These two selections of medication stops were merged to the corresponding clinical notes and resulted in a dataset of 3,000 unique clinical notes, as there were some notes linked to medication stops that reported ADRs as well as medication stops that did not report an ADR.

Research within the CCN database does not fall under the Dutch Medical Research Involving Human Subjects Act.

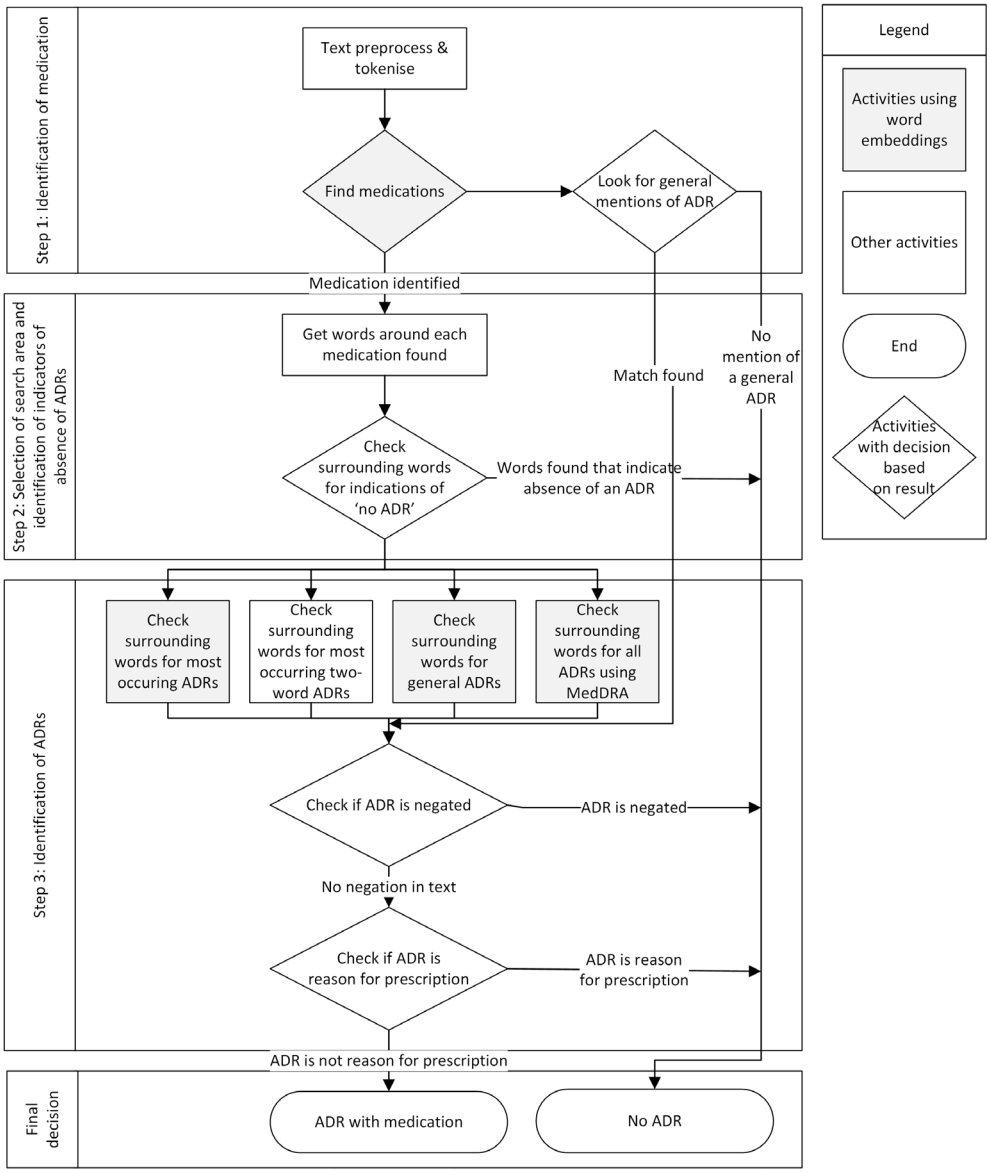


Figure 1 Overview of the different steps in the ADRIN method. ADR: adverse drug reaction.

Labelling

Two researchers independently labelled all clinical notes in the test set. Clinical notes that contained ADR information were labelled as positive. When a note was labelled positively, all words in the text that described the medication and ADR combinations were extracted. Discrepancy between labelling by the two researchers was discussed and interobserver variability was evaluated. Furthermore, a validation dataset of 3,000 unique clinical notes were labelled by one of the researchers. These notes were used for identification of thresholds for the word embedding models and for intermediate, qualitative and direct feedback.

Pre-processing clinical notes

Before applying word embedding models to the clinical notes, the text underwent multiple pre-processing steps. First, all text was converted to lowercase and unidecoded. Second, the clinical notes were tokenized with a regular expression tokenizer set to greedy tokenization for every word in the presented text Third, all numerical tokens were converted into its written form (number normalization[19]). It is assumed that this results

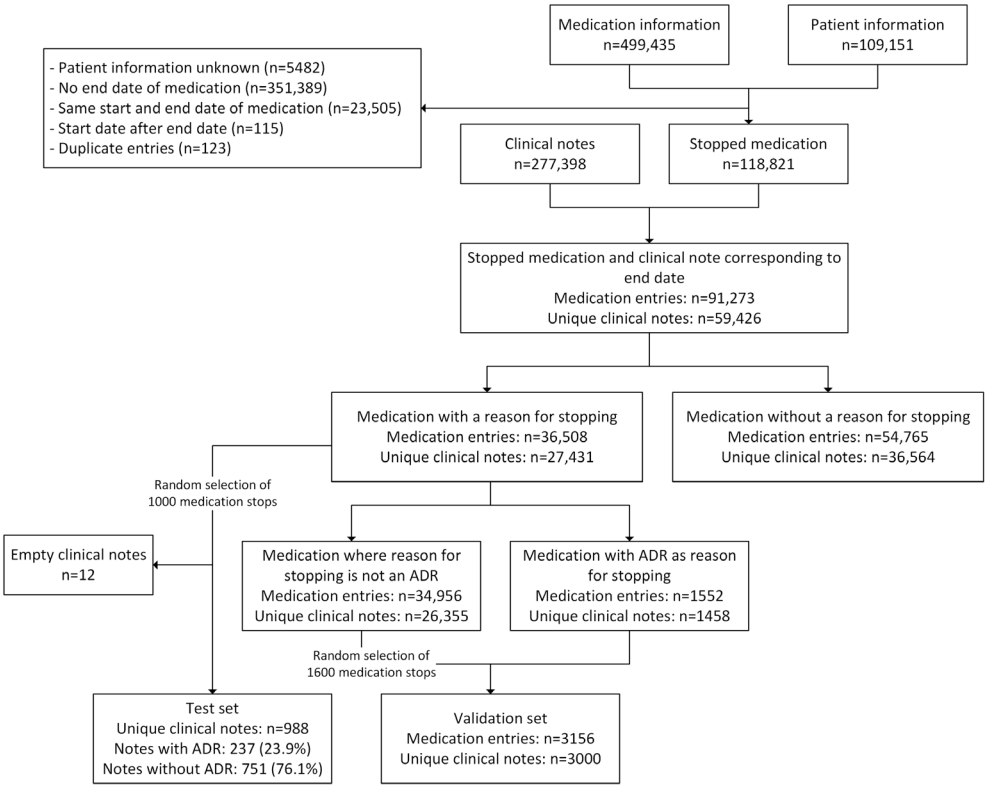


Figure 2 Flowchart of selection of clinical notes and corresponding ADR and medication

in numbers being more closely related in vector space, i.e. "16" and "18" vs "sixteen" and "eighteen". Doses were removed from the text with regular expressions. Removal of the doses was done to reduce the similarity between frequently prescribed doses and specific medication. This would otherwise contaminate the word embedding models used for identification of medication. Finally, for each token a check was performed to determine if the token was in the unigram word embedding model. If this was not the case, the word was removed from the list of tokens. An example of a text going through this process is presented in Supplementary figure 2. Preprocessing of the text was done in Python (Python Software Foundation, https://www.python.org, version 3.7.9) with the nltk package (version 3.5)[21].

Word embedding models

For the automatic identification of ADRs from text, word embedding models were developed. Two Word2Vec models, imported from the Gensim package (version 3.8.0)[22] were trained on the complete set of 277,398 clinical notes.[23] A unigram model was developed with vectors for single words. This model included all words and derived vectors that occurred more than once in the complete set of clinical notes. The second model used a combination of single words, bigrams and the derived vectors (bigram model). For the development of this model, words that occurred together more than 5 times were represented as a vector. Stop words, imported from the nltk package[21], were removed from the text. A skipgram approach was used.

Word2Vec settings were; vector size of 200 dimensions, a window of 5 words around the main word and 5 iterations of learning. Word embedding models were qualitatively evaluated through inspection of similarity between words.[24]

Identification of medication and ADRs

A list of search words was created for both medication and ADRs. The medication search list was based on different groups of cardiovascular medication (Supplementary table 8). For ADR identification, the most frequently reported ADRs (Supplementary table 7) in the discontinued medication were considered. From these ADRs a list of search words for ADR recognition was composed (Supplementary table 8).

Word embeddings were used for evaluation of the clinical note. First, the cosine similarity between each word in the clinical note and the search words for medication was calculated. A medication was identified if the cosine similarity was above a certain predefined threshold (Supplementary table 8). If no medication was found in the text, a second search was performed to identify a mention of ADRs with more general search words like 'adverse drug reaction'. If this also returned negative, the clinical note was automatically labelled as not containing an ADR (Figure 1, step 1).

Second, after identification of medication, the clinical note was searched for ADRs using

a predefined search area around the identified medication (Figure 1, step 2). This search area was restricted to prevent an increasing number of false positives and could be adjusted if it seemed too strict or too wide. This was one of the settings that was adjusted during the evaluation of the pipeline.
After this, the area was checked for 'non-ADR keywords'. These words occur right before or after the medication and indicate a medication change or extension, like 'increase' and 'double'. Therefore, these words do not indicate the presence of an ADR. List comparison was used, in which the tokenized form of the clinical note was compared with a list of words that point towards a medication change, not likely due to an ADR (Supplementary table 9).
The final step in the search for ADRs was the actual identification (Figure 1, step 3). Two sequential approaches were developed for this. The first approach included the application of the MedDRA. A selection of the lower level MedDRA (Lowest Level Terms - LLT)[17] terms was checked with text retrieval and string matching in the defined search area around the medication. In- or exclusion of the MedDRA was one of the settings adjusted during the evaluation of the pipeline.
The second approach for identification of ADRs was the use of unigram and bigram word embedding models. For each word in the search area the cosine similarity with the search words for ADRs was computed (Supplementary table 8). In case this similarity was above the predefined threshold, the word was identified as an ADR. Threshold setting was done with a grid search. Visual inspection of the graphical representation of the number of correct matches for a specific word (Supplementary Figure 1) and evaluation of the included words after inspection of the list of most similar words (e.g. Table 3) resulted in the setting of the thresholds. For example, in case of a specific medication, the threshold was set in a way that spelling mistakes and closely related medication were selected, but not words that were related to a significant other medication group or words that did not describe medication, a certain disease or condition. For this analysis the validation dataset was used. This is explained in more detail in the supplemental materials.

Pipeline versions and tasks

The pipeline was developed to execute four different tasks; a binary classification whether the clinical note contained an ADR or not (A), the exact extraction of the medication and corresponding ADR (B), and the extraction of the medication that causes an ADR (C) and the ADR (D) individually.
Multiple settings were changed during the analysis to evaluate the performance of the predefined tasks of different experimental designs of the pipeline; in- or exclusion of MedDRA for ADR identification, inclusion or neglect of punctuation for demarcation of the search area and size of the search area. Table 1 gives an overview of the different

settings that were evaluated in the versions of th
e pipeline. Analysis of the pipeline was done in Python (Python Software Foundation, https://www.python.org, version 3.7.9).

Performance metrics

The pipeline was evaluated on the test set of 988 labelled clinical notes. Different metrics were calculated to assess the performance of different versions of the pipeline. The metrics that were calculated included; accuracy and balanced accuracy, sensitivity, specificity, precision/positive predictive value, negative predictive value, recall, F_1-score, detection rate and detection prevalence. An elaborate overview of the performance metrics and the evaluation process can be found in Supplementary table 1 and Supplementary table 2 to 6, respectively. Evaluation of the outcome was done with the R programming language (R Foundation for Statistical Computing, https://ww.R-project.org , version 4.0.2) and with RStudio (RStudio: Integrated Development Environment for R, http://www.rstudio.com/, version 1.3.1093). The caret package was used for evaluation (version 6.0-86)[25].

RESULTS

Dataset

Information on the complete dataset for word embeddings models, validation set and test set is described in Table 2. Characteristics of the included free text are the informal writing style, use of abbreviations and relatively short length of the text. The supplemental materials contain four different translated examples of clinical notes, as shown in Supplementary table 2.

Word embedding models

Several search terms of the prototype have been independently reviewed in the word embedding models to evaluate the performance of the word embedding models. Table 3 shows these key words and the five most similar words. It was noted that in case the search word was a specific group of medication (e.g. betablocker), also other groups of

Table 1 Settings of the pipeline features of the different computational experiments.

Version	Words in search area	Considering punctuation	Version without MedDRA
1A	All	Yes	1B
2A	All	No	2B
3A	10	Yes	3B
4A	10	No	4B
5A	5	Yes	5B
6A	5	No	6B

medications were identified (e.g. 'diltiazem' in case of search word 'beta-blocker'). As the identified word is used for the analysis, and not the search word, this has no consequences for the analysis.
In the training of the word embedding models, free text from clinical notes was used. This is domain-specific data, which can give improve the embedding of domain-specific words. An illustrative example is the word embedding of 'red'. In our word embedding models, trained specifically on medical text, 'red' is closely associated with 'itching', 'swollen', 'irritated' and 'colourings', whereas in word embeddings on general text, 'red' would be associated with other colours.

Interobserver variability

A test set (n=988 clinical notes) was manually labelled by two independent researchers and was used for the evaluation of the pipeline. During this process, 908 notes were identically labelled. This resulted in an interobserver variability for the binary presence of an ADR of 91%. Regarding the literal extraction of the ADR and the medication, there were 215 instances (21.7%) where the result differed between the researchers. This was mostly due to a difference in taking adjectives or adverbs into account or a different interpretation of the clinical note. As the pipeline is trained on one-word and two-word ADRs it was decided that these words were not taken into account.
Manual labelling of the 988 clinical notes in the test set resulted in 237 (23.9%) notes that were binary classified as containing an ADR. These notes contained in total 392 combinations of a triggered ADRs and corresponding medication.

Evaluation of pipeline

Figures 3 and 4 show the performance of the pipeline on the different metrics and for the different tasks. Supplementary table 10 shows the values for true and false negatives and true and false positives per version and per task. The task for binary classification achieved the highest accuracy, varying from 0.70 to 0.84 (Figure 3A). However, as this was

Table 2 Characteristics of selected clinical notes for development of word embedding models, validation set and the test set.

Variable	Word embedding models	Validation set	Test set
Language	Dutch	Dutch	Dutch
Number of unique records	277.398	2999	988
Unique patients	108,940	2,690	955
Number of unique tokens	96,086	9,297	5,464
Average number of tokens per record	54	53	53
Female sex of individuals, n (%)	56,527 (51)	1,320 (49)	459 (47)

the easiest task, the accuracy of the pipeline on the exact extraction of medication and ADR together was much lower, varying from 0.23 to 0.64 (Figure 3D).
If we look at the specific settings of the different pipelines, the results show that the addition of the MedDRA to the pipeline did not lead to an increase in the performance of the pipeline (Figure 4A-4D). Overall, the inclusion of punctuation led to a better performance than transcending sentences (version 1, 3, and 5), and a search area of 5 words seemed to lead to the best results overall (version 5 and 6).
The negative predictive value - the chance that no ADR was present when the pipeline did not produce an ADR- was approximately the same per task (0.69-0.91) for all versions of the pipeline. However, the positive predictive value (i.e. the chance that, when the pipeline reported an ADR, it was in fact reported in the clinical notes) varied much more per version (Figure 3 and 4) and varied between 0.071 and 0.71. This could be explained by the proportion of false negatives. The proportion of false negatives did not vary much

Table 3 Selection of results from the word embedding models; ADRs and medication search words and a selection of the most relevant similar words, where spelling mistakes are excluded. Similarity is based on the cosine similarity.

Key word	Most similar words in Dutch (English, cosine similarity)			
Pijn op de borst (chest pain)	Druk op de borst (chest pressure, 0.80)	Kramp op de borst (chest crampings, 0.70)	Pijn in de armen (pain in the arms, 0.68)	Retrosternale pijn (pain retrosternal, 0.67)
Verminderde conditie (decreased condition)	Afname conditie (decreasing stamina, 0.63)	Conditieverlies (loss of condition, 0.63)	Verminderde inspanningstolerantie (decreased exercise tolerance, 0.62)	Overmatig transpireren (excessive sweating, 0.62)
Oedeem (edema)	Perifeer (peripheral edema, 0.81)	Enkeloedeem (ankle edema, 0.80)	Pitting (pitting, edema 0.80)	Enkels (ankles edema, 0.75)
Hoesten (coughing)	Sputum (sputum, 0.75)	Slijm (mucus, 0.71)	Hoestklachten (coughing complaints, 0.70)	Kuchen (to cough, 0.70)
Duizelig (dizzyness)	Zweterig (sweaty, 0.73)	Misselijk (nauseous, 0.71)	Zweverig (floaty, 0.70)	Draaierig (dizzy, 0.69)
Statine (statin)	Simvastatine (Simvastatin, 0.80)	Pravastatine (Pravastatin, 0.76)	Crestor (Rosuvastatin, 0.75)	Atorvastatine (Atorvastatin, 0.74)
Betablokker (Beta-blocker)	Metoprolol (0.74)	Atenolol (0.71)	Diltiazem (0.66)	Bisoprolol (0.65)
Antistolling	Acenocoumarol (acenocoumarin, 0.80)	Anticoagulantia (Anticoagulants, 0.78)	NOAC (novel oral anticoagulant, 0.77)	Fenprocoumon (phenprocoumon, 0.74)
Amlodipine	Nifedipine (0.85)	Lisinopril (0.82)	Barnidipine (0.81)	Enalapril (0.79)

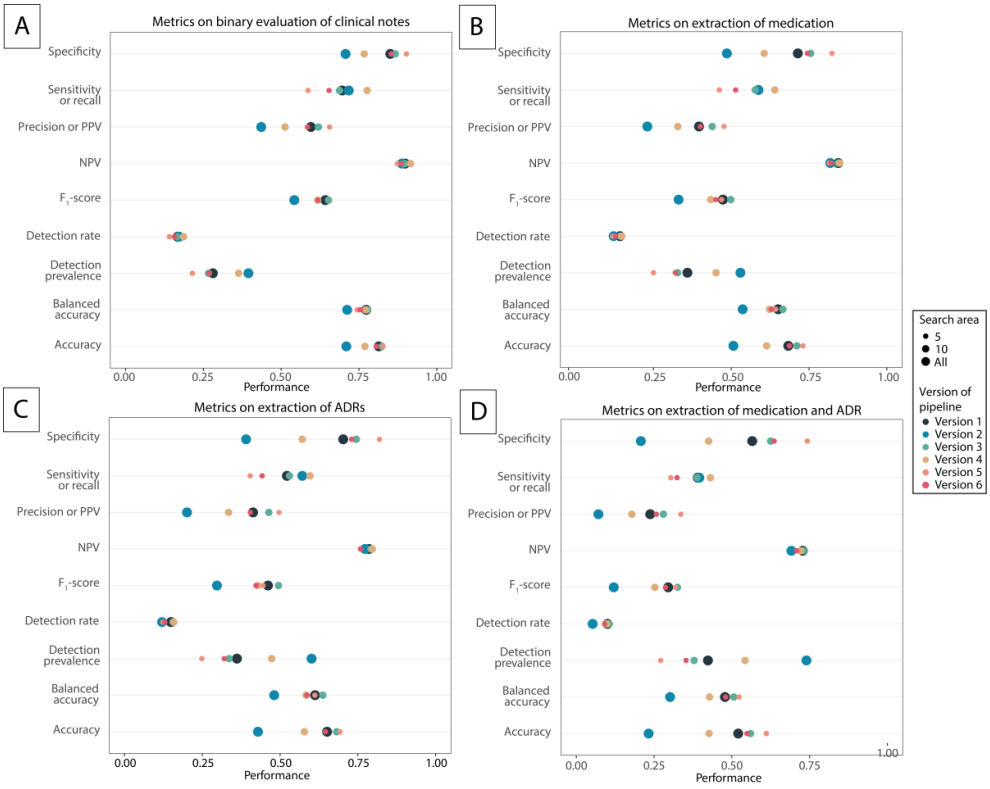


Figure 3 Performance of different experimental versions of the pipeline with the inclusion of the MedDRA on the different tasks (A: binary evaluation, B: medication identification, C: ADR identification, D: medication and ADR + adverse drug reaction identification). ADR: adverse drug reaction; MedDRA: Medical Dictionary for Regulatory Activities; NPV: negative predictive value; PPV: positive predictive value.

per version of the pipeline for a given task. However, the proportion of false positives had much more variety, caused by a change in the search area and the inclusion or exclusion of punctuation, which led to more ADRs found with a specific medication.

The optimal version of the pipeline depends on the task for which the pipeline is used. If the task is to select notes based on whether they contain ADRs, the results of the binary classification task (task 1) are most relevant. For this task, version 3B (i.e., no use of the MedDRA, search area of 10 words, and considering punctuation) generated the highest accuracy (0.84) and F1 score (0.67). In this case, 8.1% (80/988) of notes were classified as false negatives, indicating that 8.1% (80/988) of notes would not be selected when looking for ADRs. The most optimal version based on accuracy for identification of medication, ADRs, and ADRs and medication combined was version 5B, with an accuracy for the different tasks of 0.75, 0.72, and 0.64, respectively. Version 3B was the optimal version when emphasis was on the F1 score, with scores of 0.52, 0.52, and 0.35 for identification

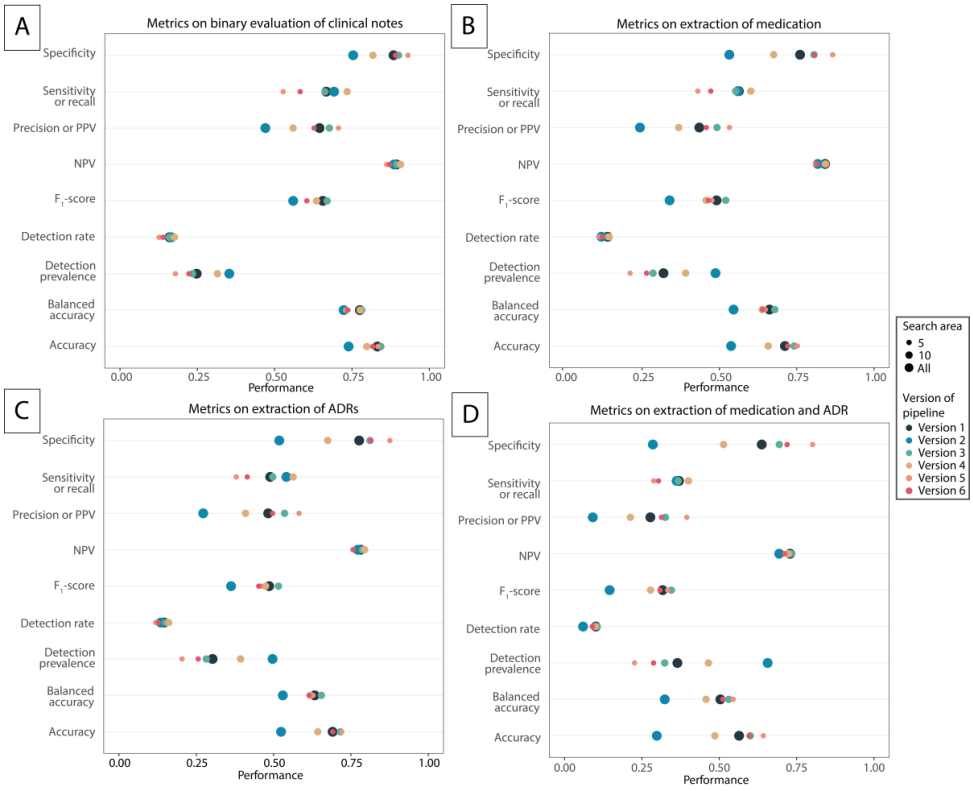


Figure 4 Performance of different experimental versions of the pipeline without the use of the MedDRA on the different tasks (A: binary evaluation, B: medication identification, C: ADR identification, D: medication and ADR + adverse drug reaction identification). ADR: adverse drug reaction; MedDRA: Medical Dictionary for Regulatory Activities; NPV: negative predictive value; PPV: positive predictive value.

of medication, ADRs, and medication and ADRs combined, respectively. During the evaluation of the notes in the test set, the prototype incorporating the MedDRA required approximately 70 minutes to generate an outcome for all notes, whereas the versions without the MedDRA took approximately 14 seconds.

DISCUSSION

In this work, the Adverse Drug Reactions Identification in clinical Notes (ADRIN) method and a corresponding prototype were developed. The method was evaluated on a subset of clinical notes. Different versions of the prototype lead to differing results on the various tasks. The optimal version of the pipeline depends on the task and the trade-off being made; is it more valuable to find as many medication and ADR-combinations as possible, or find less ADRs, but also make less mistakes? If the goal is the first, a larger search area is better. However, even with the entire note as search area, at least 8% of all

medication and ADR combinations were missed. When you want to be more accurate, a smaller search area is preferred and punctuation should be taken into account. This reduces the number of false positives generated, which results in increased accuracy and F1-score.

Surprisingly, the versions with incorporation of the MedDRA dictionary performed worse on most tasks than the same version without the MedDRA. The negative effect on the performance by the MedDRA is due to the large increase in false positives it generated. This is caused by string matching with the MedDRA dictionary, leading to more identifications than the specific set of frequently occurring ADRs defined by the predefined search words. Yet, incorporation of the MedDRA can possibly lead to an improved uptake of rare ADRs, but this has not been evaluated in more detail. Furthermore, misspelled ADRs are not recognized by the MedDRA search, creating added value of the incorporation of word embedding models. Moreover, implementation of the MedDRA in the prototype significantly increased execution time, a significant attribute if real-time evaluation of clinical notes is required.

Illustrative for the underreporting of ADRs is that in 54,765 (60.1%) of the discontinued medication entries no reason was reported for ending of the medication in the registration of a patient's medications. Yet, there were 36,564 (61.5%) clinical notes matched to these medication entries, which illustrates the potential additional value of clinical notes in unravelling ADRs in this dataset.

When we put the presented results in light of the ongoing developments of ADR extraction from clinical notes, we see that the performance of our pipeline is similar to other presented pipelines. First of all, most publications focus on the automatic extraction of ADRs, ADEs or adverse events[26–29], whereas our study identifies the combination of medication and triggered ADR. Another publication that identified both ADR and medication showed increased performance, with F-scores for drug, ADR and combination drug and ADR of respectively, 0.930, 0.887 and 0.534[30], versus performance that we showed of 0.52, 0.51 and 0.34. When comparing methodologies, our method dominantly relies upon internal information and similarity from word embeddings, whereas Tang et al. use external reference sources for development of their dictionaries, which is the case in most studies.[29,31] The use of word embeddings increases the identification of spelling mistakes in medication and ADRs, brand names and synonyms. Yet, in our methodology, there was also an increased number of false positives.

Word embedding models can thus be used for the identification of spelling mistakes and brand names of medication. However, for the identification of synonyms, the use case must be critically evaluated. It was shown that words that indicate what is done with a specific prescription, e.g., 'to lower' and 'to increase', are considered similar by the word embedding models. It is therefore not suitable to use word embedding models

for identification of 'non-ADR' keywords, which has been solved with string matching in the ADRIN method. The use of domain-specific word embedding models is not new and limited to ADR identification, but is increasingly used in evaluation of clinical notes, e.g. in ICD-10 classification[15] and anonymization[32].

Second, publications for identification of ADRs in the English language are numerous, using different methods, like GATE NLP[31], trigger words[27] or trigger phrases[28]. Regarding foreign languages, the field is maturing. Methods developed for the English language can, in some cases, be transferred to other languages. However, the effort that has to be put into this depends on complexity of the task and the level of text interpretation.[33] For example, a study in Danish clinical notes obtained better performance (recall of 0.75 vs. 0.59 respectively) for sole ADR identification. This study misses approximately one fourth of all possible ADRs, whereas our optimal performance misses approximately 40%. However, this pipeline included manual dictionary selection and more rule-based filters in the model.[29]

We have chosen to use the presence of a mention of medication in the clinical note as the starting point for identification of an ADR. However, this might result in experienced ADRs that are possibly missed. The performance of the pipeline might benefit from removal of the identification of medication and, for example, coupling with structured medication prescriptions to obtain information about medication use. However, the end-user should be aware that this also might increase the number of false positives, as the presence of an ADR is no longer limited by the presence of medication.

Limitations that have been identified during the evaluation of the method and prototype are primarily related to missed ADRs from the clinical free text, even when the entire clinical note is used for analysis. This problem can be solved by lowering the identifying threshold, but this would also lead to a potentially large increase in false positives. The use of machine and deep learning models can improve the performance of the ADRIN method. However, a large dataset of labelled clinical notes is required to train machine and deep learning models, which was unavailable during development of this model.

An overall limitation of the prototype is the direct translatability to other languages. The word embedding models have been specifically trained on Dutch clinical notes. To implement this method in clinical notes in a different language search terms for word embedding functions must be translated to the new language. Moreover, word embedding models have to be trained with notes in the specific language, before applying the developed method. Therefore, a large amount of clinical free text notes is required. Due to ethical and privacy constraints, this can be hard to acquire. Yet, it is technically possible to test and validate the ADRIN method in other languages by translation of search words and negations and after training of word embedding models with the specific language.

To conclude, the ADRIN method and prototype are effective in recognizing adverse drug

reactions in Dutch clinical notes. Surprisingly, incorporation of the MedDRA did not result in improved identification on top of word embedding models. However, not all versions of the prototype are equally accurate. Different parameter settings can be chosen for the prototype to optimize the task of the model. In a future stage, incorporation of a pipeline in the EHR environment can lead to automatic identification and registration of ADRs. This saves precious time of the physician and decreases the previously mentioned underreporting of ADRs in clinical care, increasing our knowledge about ADRs, which might benefit the patient in the end.

REFERENCES

1. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions: A systematic review. *Drug Saf*. 2006;29(5):385-394. doi:10.2165/00002018-200932010-00002
2. Seruga B, Templeton AJ, Badillo FEV, Ocana A, Amir E, Tannock IF. Under-reporting of harm in clinical trials. *Lancet Oncol*. 2016;17(5):e209-e219. doi:10.1016/S1470-2045(16)00152-2
3. Leening MJG, Heeringa J, Deckers JW, et al. Healthy volunteer effect and cardiovascular risk. *Epidemiology*. 2014;25(3):470-471. doi:10.1097/EDE.0000000000000091
4. de Vries ST, Denig P, Ekhart C, et al. Sex differences in adverse drug reactions reported to the National Pharmacovigilance Centre in the Netherlands: An explorative observational study. *Br J Clin Pharmacol*. 2019;85(7):1507-1515. doi:10.1111/bcp.13923
5. Kongkaew C, Noyce PR, Ashcroft DM. Hospital Admissions Associated with Adverse Drug Reactions: a systematic review of prospective observational studies. *Ann Pharmacother*. 2008;42:1017-1025.
6. Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother*. 2013;4(Supplement 1). doi:10.4103/0976-500X.120957
7. Postigo R, Brosch S, Slattery J, et al. EudraVigilance Medicines Safety Database: Publicly Accessible Data for Research and Public Health Protection. *Drug Saf*. 2018;41(7):665-675. doi:10.1007/s40264-018-0647-1
8. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309(13):1351-1352. doi:10.1001/jama.2013.393
9. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: Systematic review. *J Med Internet Res*. 2019;21(5):1-18. doi:10.2196/12239
10. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol*. 2020;145(2):463-469. doi:10.1016/j.jaci.2019.12.897
11. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. *J Biomed Inform*. 2019;100S:100057. doi:10.1016/j.yjbinx.2019.100057
12. Zhao M, Masino AJ, Yang CC. A Framework for Developing and Evaluating Word Embeddings of Drug-named Entity. In: *Proceedings of the BioNLP 218 Workshop*; 2018:156-160. doi:10.18653/v1/w18-2319
13. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87(July):12-20. doi:10.1016/j.jbi.2018.09.008
14. Banerjee I, Madhavan S, Goldman RE, Rubin DL. Intelligent word embeddings of free-text radiology reports. *arXiv*. Published online 2017:411-420.
15. Sammani A, Bagheri A, van der Heijden PGM, et al. Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks. *npj Digit Med*. 2021;4(1). doi:10.1038/s41746-021-00404-9
16. Dai HJ, Su CH, Wu CS. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. *J Am Med Informatics Assoc*. 2020;27(1):47-55. doi:10.1093/jamia/ocz120
17. Brown EG, Wood L, Wood S. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Saf*. 1999;20(September 1998):109-117.
18. Bots SH, Siegersma KR, Onland-Moret NC, et al. Routine clinical care data from thir-

teen cardiac outpatient clinics: design of the Cardiology Centers of the Netherlands (CCN) database. *BMC Cardiovasc Disord.* 2021;21(1):1-9. doi:10.1186/s12872-021-02020-7
19. Menger V, Scheepers F, van Wijk LM, Spruit M. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telemat Informatics*. 2018;35(4):727-736. doi:10.1016/j.tele.2017.08.002
20. Sproat R, Black AW, Chen S, Kumar S, Ostendorf M, Richards C. Normalization of non-standard words. *Comput Speech Lang*. 2001;15(3):287-333. doi:10.1006/csla.2001.0169
21. Bird S. NLTK: The Natural Language Toolkit. In: *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions.* ; 2006:69-72. doi:10.3115/1225403.1225421
22. Rehurek R, Sokja P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA*; 2010:45-50. http://is.muni.cz/publication/884893/en
23. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *1st Int Conf Learn Represent ICLR 2013 - Work Track Proc.* Published online 2013:1-12.
24. Wang B, Wang A, Chen F, Wang Y, Kuo CCJ. Evaluating word embedding models: Methods and experimental results. *APSIPA Trans Signal Inf Process.* 2019;8:1-13. doi:10.1017/ATSIP.2019.12
25. Kuhn M. The caret Package. Published online 2009.
26. Honigman B, Lee J, Rothschild J, et al. Using computerized data to identify adverse drug events in outpatients. *J Am Med Inf Assoc*. 2001;8(3):254-266.
27. Murff HJ, Forster A j., Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically sreening discharge summaries for adverse medical events. *J Am Med Informatics Assoc*. 2003;10(4):339-351. doi:10.1197/jamia.M1201.Affiliations
28. Cantor MN, Feldman HJ, Triola MM. Using trigger phrases to detect adverse drug reactions in ambulatory care notes. *Qual Saf Heal Care*. 2007;16(2):132-134. doi:10.1136/qshc.2006.020073
29. Eriksson R, Jensen PB, Frankild S, Jensen LJ, Brunak S. Dictionary construction and identification of possible adverse drug events in danish clinical narrative text. *J Am Med Informatics Assoc*. 2013;20(5):947-953. doi:10.1136/amiajnl-2013-001708
30. Tang Y, Yang J, Ang PS, et al. Detecting adverse drug reactions in discharge summaries of electronic medical records using Readpeer. *Int J Med Inform*. 2019;128(November 2018):62-70. doi:10.1016/j.ijmedinf.2019.04.017
31. Iqbal E, Mallah R, Jackson RG, et al. Identification of adverse drug events from free text electronic patient records and information in a large mental health case register. *PLoS One*. 2015;10(8):1-14. doi:10.1371/journal.pone.0134208
32. Abdalla M, Abdalla M, Rudzicz F, Hirst G. Using word embeddings to improve the privacy of clinical notes. *J Am Med Informatics Assoc*. 2020;27(6):901-907. doi:10.1093/jamia/ocaa038
33. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *J Biomed Semantics*. 2018;9(1):1-13. doi:10.1186/s13326-018-0179-8

SUPPLEMENTARY MATERIALS

Evaluation of pipeline

The different versions of the pipeline are evaluated on different tasks. To evaluate the performance of each version of the pipeline on different tasks, different metrics and their calculations are shown in Supplementary table 1.

Supplementary table 2 gives four different examples of the result of manual labelling and the result of a version of the pipeline. In Supplementary table 3, Supplementary table 4, Supplementary table 5 and Supplementary table 6, the number of true positives (TN), false positives (FP), true negatives (TN) and false negatives (FN) is shown per example.

Supplementary table 1 Overview of used metrics.

Metrics	Calculation	Explanation
Accuracy	(TP + TN)/(TN + TP + FN + FP)	Proportion of samples that have been correctly classified.
Sensitivity/Recall	TP/(TP + FN)	Proportion of true positives that are classified correctly, given the number of actual positives.
Specificity	TN/(TN + FP)	Proportion of true negatives that are classified correctly, given the number of actual negatives.
Precision/Positive predictive value (PPV)	TP/(TP + FP)	Proportion of true positives that are classified correctly, given the number of positive classifications.
Negative predictive value (NPV)	TN/(TN + FN)	Proportion of true negatives that are classified correctly, given the number of negative classifications.
F1-score	2*((precision*recall)/(precision + recall))	Balance between precision and recall.
Detection rate	TP/(TN + FP + FN + FP)	Proportion of classifications that are correctly positively classified.
Detection prevalence	(TN + FP)/(TN + TP + FN + FP)	Proportion of classifications that are positively classified
Balanced accuracy	(Sensitivity + Specificity)/2	Average of sensitivity and specificity.

TP: true positives, TN: true negatives, FP: false positives, FN: false negatives

Supplementary table 2 Four examples of consult notes, the result of manual labelling and a possible result of the pipeline.

Text	Manual labelling	Manual label	Result Pipeline	Pipeline label
Patient has developed evident hypertension with signs of clvh/strain, c.q. diastolic lv dysfunction. On the other hand, no signs for increased r pressure or dilated r structures. Pat was convinced to change antihypertensive medication.	[]	0	[]	0
This is a patient with elaborate cardiac history, at this moment hypertension, ace inhibitor was stopped. Besides, statin was halved.	[]	0	[ace, hypertension], [inhibitor, hypertension]	1
<Person> because hefty rectal bleeding with known colitis. Ufn stop ascal.	[ascal, rectal bleeding]	1	[]	0
Monitoring after ablation, no complaints (probably atrial fibrillation once, has lost more than 10 kilograms) is doing well. Echocardiogram and stress test did nothing details. Decrease metoprolol. Control 3 months. C/ state after uncomplicated isolation of lung veins icw atrial fibrillation without complaints amiodaron decreased normal stress ecg normal echocardiogram hypertriglyceridaemia b/ decrease metoprolol wrt hypertriglyceridaemia (temporarily) stop statin/start modalim co 3 months	[statin, hypertriglyceridaemia]	1	[metoprolol, hypertriglyceridaemia], [metoprolol, atrial fibrilation]	1

Task A: Binary evaluation

Supplementary table 3 Four examples of consult notes and the evaluation of binary labelling for the presence or absence of ADR information.

Text	Manual	Pipeline	Class
Patient has developed evident hypertension with signs of clvh/strain, c.q. diastolic lv dysfunction. On the other hand, no signs for increased r pressure or dilated r structures. Pat was convinced to change antihypertensive medication.	0	0	TN
This is a patient with elaborate cardiac history, at this moment hypertension, ace inhibitor was stopped. Besides, statin was halved.	0	1	FP
<Person> because hefty rectal bleeding with known colitis. Ufn stop ascal.	1	0	FN
Monitoring after ablation, no complaints (probably atrial fibrillation once, has lost more than 10 kilograms) is doing well. Echocardiogram and stress test did nothing details. Decrease metoprolol. Control 3 months. C/ state after uncomplicated isolation of lung veins icw atrial fibrillation without complaints amiodaron decreased normal stress ecg normal echocardiogram hypertriglyceridaemia b/ decrease metoprolol wrt hypertriglyceridaemia (temporarily) stop statin/start modalim co 3 months	1	1	TN

Task B: Evaluation of ADR

Supplementary table 4 Four examples of consult notes and the evaluation of the outcome of the pipeline on the extraction of ADRs.

Text	Manual	Pipeline	Class
Patient has developed evident hypertension with signs of clvh/strain, c.q. diastolic lv dysfunction. On the other hand, no signs for increased r pressure or dilated r structures. Pat was convinced to change antihypertensive medication.	[]	[]	TN
This is a patient with elaborate cardiac history, at this moment hypertension, ace inhibitor was stopped. Besides, statin was halved.	[]	[hypertension], [hypertension]	FP (1, because 1 unique ADR)
<Person> because hefty rectal bleeding with known colitis. Ufn stop ascal.	[rectal bleeding]	[]	FN (1)
Monitoring after ablation, no complaints (probably atrial fibrillation once, has lost more than 10 kilograms) is doing well. Echocardiogram and stress test did nothing details. Decrease metoprolol. Control 3 months. C/ state after uncomplicated isolation of lung veins icw atrial fibrillation without complaints amiodaron decreased normal stress ecg normal echocardiogram hypertriglyceridaemia b/ decrease metoprolol wrt hypertriglyceridaemia (temporarily) stop statin/start modalim co 3 months	[hypertriglyceridaemia]	[hypertriglyceridaemia], [atrial fibrilation]	TP (1) FP (1)

Task C: Medication evaluation

Supplementary table 5 Four examples of consult notes and the evaluation of the outcome of the pipeline on the extraction of medication that caused the ADR.

Text	Manual	Pipeline	Class
Patient has developed evident hypertension with signs of clvh/ strain, c.q. diastolic lv dysfunction. On the other hand, no signs for increased r pressure or dilated r structures. Pat was convinced to change antihypertensive medication.	[]	[]	TN
This is a patient with elaborate cardiac history, at this moment hypertension, ace inhibitor was stopped. Besides, statin was halved.	[]	[ace], [inhibitor]	FP (2)
<Person> because hefty rectal bleeding with known colitis. Ufn stop ascal.	[ascal]	[]	FN (1)
Monitoring after ablation, no complaints (probably atrial fibrillation once, has lost more than 10 kilograms) is doing well. Echocardiogram and stress test did nothing details. Decrease metoprolol. Control 3 months. C/ state after uncomplicated isolation of lung veins icw atrial fibrillation without complaints amiodaron decreased normal stress ecg normal echocardiogram hypertriglyceridaemia b/ decrease metoprolol wrt hypertriglyceridaemia (temporarily) stop statin/start modalim co 3 months	[statin]	[metoprolol]	FP (1) FN (1)

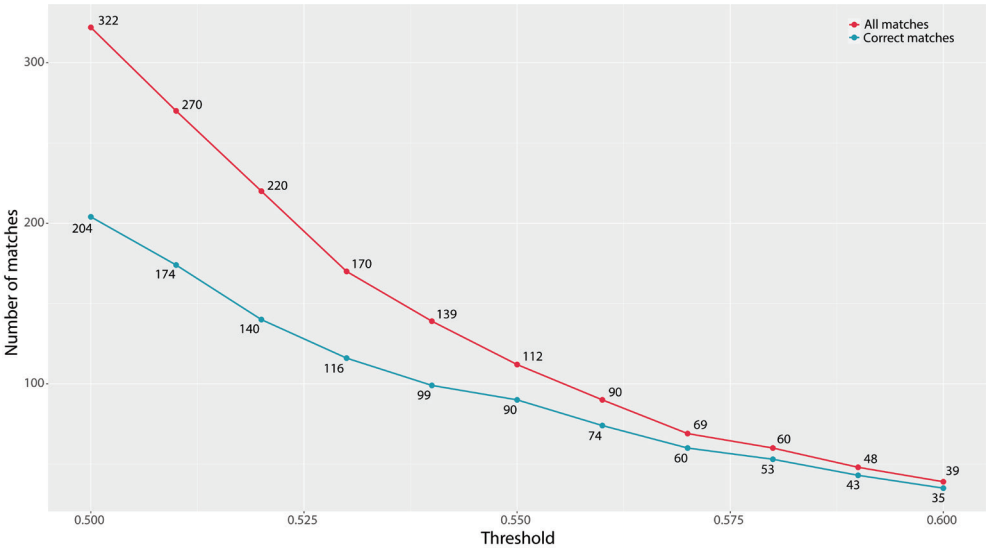
Task D: Evaluation of medication and ADR

Supplementary table 6 Four examples of consult notes and the evaluation of the outcome of the pipeline on the extraction of the medication and ADR combination.

Text	Manual	Pipeline	Class
Patient has developed evident hypertension with signs of clvh/strain, c.q. diastolic lv dysfunction. On the other hand, no signs for increased r pressure or dilated r structures. Pat was convinced to change antihypertensive medication.	[]	[]	TN
This is a patient with elaborate cardiac history, at this moment hypertension, ace inhibitor was stopped. Besides, statin was halved.	[]	[ace, hypertension], [inhibitor, hypertension]	FP (2)
<Person> because hefty rectal bleeding with known colitis. Ufn stop ascal.	[ascal, rectal bleeding]	[]	FN (1)
Monitoring after ablation, no complaints (probably atrial fibrillation once, has lost more than 10 kilograms) is doing well. Echocardiogram and stress test did nothing details. Decrease metoprolol. Control 3 months. C/ state after uncomplicated isolation of lung veins icw atrial fibrillation without complaints amiodaron decreased normal stress ecg normal echocardiogram hypertriglyceridaemia b/ decrease metoprolol wrt hypertriglyceridaemia (temporarily) stop statin/start modalim co 3 months	[statin, hypertriglyceridaemia]	[metoprolol, hypertriglyceridaemia], [metoprolol, atrial fibrilation]	FP (2) FN (1)

Cosine similarity in word embedding models

The cosine similarity between search words and the words in the search area of the clinical note was used to extract medications and ADRs from text. Thresholds for cosine similarity were defined based on the trade-off between the number of identified ADRs and correct matches. This analysis was performed on the validation set. For each search word, a grid search was performed with thresholds between 0.50 and 0.60 to determine the optimal threshold value. A threshold was determined in a way that it captures a significant number of words, without generating too many false positives and are based on visual inspection of figures such as Supplementary Figure 1 and evaluation of the most similar words as is displayed for a specific selection in Table 3. False positives are generated when the threshold is set at a value that is too low. Furthermore, different search words had overlap in their identified matches. Therefore, the chosen threshold could be set to a higher value, because another synonym of the word is already captured by another search word. This reduces the number of false positives. An example of threshold evaluation is shown in Supplementary figure 1 for the word 'dizziness. The selected threshold was set to 0.58.



Supplementary figure 1 Grid search for selection of threshold to evaluate the number of (correct) matches for the search word 'dizziness'.

Supplementary table 8 Search words - Dutch and (English) - for the identification of ADRs, medication and the predefined cosine similarity threshold.

ADR Search Words	Threshold
Myalgie (myalgia)	0.50
Nierfunctiestoornis (kidney dysfunction)	0.60
Kriebelhoest (cough)	0.55
Oedeem (edema)	0.60
Hypotensie (hypotension)	0.57
Aritmie (arrythmia)	0.50
Moeheid (fatigue)	0.55
Duizeligheid (dizzyness)	0.58
Hoofdpijn (headache)	0.56
Hematome (haematoma)	0.54
Atriumfibrileren (atrial fibrillation)	0.59
pijn_op_de_borst (chest pain)	0.60
lage_rr (low blood pressure)	0.51
verminderde_conditie (decreased stamina)	0.60
dikke_enkels (swollen ankles)	0.60
hoge_bloeddruk (high blood pressure)	0.58
hoge_hr (high heart rate)	0.58
chonotrope_incompetentie (chronotropic incompetence)	0.70
traag_sinusritme (slow sinusrhythm)	0.77
qrs_verbreding (qrs elongation)	0.72
depressieve_gevoelens (depressed feelings)	0.75
laag_kalium (low potassium)	0.65
pijnlijk_gevoel (painful feelings)	0.80

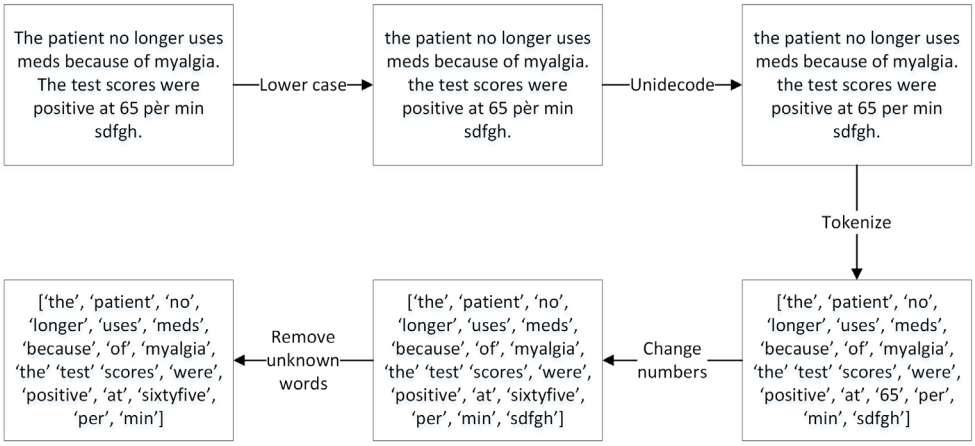
Medication Search Words	Threshold
Acenocoumarol (Acenocoumarol)	0.68
Amiodarone (Amiodaron)	0.64
Amlodipine (Amlodipine)	0.67
Aspirine (Aspirin)	0.65
Clopidogrel (Clopidogrel)	0.63
Gemfibrozil (Gemfibrozil)	0.76
Hydrochloorthiazide (Hydrochlorothiazide)	0.66
Medicatie (Medication)	0.61
Metoprolol (Metoprolol)	0.70
Beta (beta)	0.63
Nitroglycerine (Nitroglycerin)	0.65
Perindopril (Perindopril)	0.68
Statine (Statin)	0.66
Tamsulosine (Tamsolusin)	0.75
Valsartan (Valsartan)	0.67

Supplementary table 7 Frequency of diferent ADRs that were manually extracted from the free text of the reason for discontinuation of a prescription.

ADR	Number of mentions, n
Myalgia	157
Oedema	84
Kidney Dysfunction	72
Cough	67
Arrhythmia	49
Hypotension	48
Fatigue	48
Dizziness	44
Headache	30
Stomach pain	27

Supplementary table 9 Words that indicate a change or the initiation of medication. These words stop further evaluation of ADRs and medication in the text.

Not-ADR keywords before Medication	Not-ADR keywords after medication
Verdubbeling (double)	Verhoogd (increased)
Verhoogde (increased)	Verdubbeld (doubled)
Dosisverhoging (increased dose)	Verdubbeling (double)
Begonnen (started)	Gestart (started)
Start (start)	Verhogen (to increase)
Ophogen (increased)	Opgehoogd (increased)
Hoogde (increased)	Uitgebreid (expanded)
Hoog (increase)	Toegevoegd (Added)
	Begonnen (started)
	Ophogen (to increase)
	Hoogde (increased)



Supplementary figure 2 Example of the preprocessing that was performed on the clinical notes.

Supplementary table 10 Overview of confusion matrix for all versions of the pipeline and all tasks.

Version	Task	TN	FN	FP	TP
version 1A	Binary	639	72	112	165
	ADR	639	174	270	190
	Medication	639	118	256	168
	ADR+Medication	639	239	491	153
version 1B	Binary	664	79	87	158
	ADR	664	186	191	178
	Medication	664	125	209	161
	ADR+Medication	664	247	378	145
version 2A	Binary	531	67	220	170
	ADR	531	156	828	208
	Medication	531	118	562	168
	ADR+Medication	531	237	2028	155
version 2B	Binary	566	73	185	164
	ADR	566	167	526	197
	Medication	566	126	498	160
	ADR+Medication	566	250	1417	142
version 3A	Binary	651	74	100	163
	ADR	651	171	223	193
	Medication	651	121	211	165
	ADR+Medication	651	240	391	152
version 3B	Binary	676	80	75	157
	ADR	676	183	157	181
	Medication	676	128	163	158
	ADR+Medication	676	248	298	144
version 4A	Binary	576	53	175	184
	ADR	576	147	432	217
	Medication	576	103	375	183
	ADR+Medication	576	223	778	169
version 4B	Binary	614	63	137	174
	ADR	614	159	296	205
	Medication	614	114	295	172
	ADR+Medication	614	235	581	157
version 5A	Binary	678	98	73	139
	ADR	678	217	149	147
	Medication	678	154	145	132
	ADR+Medication	678	273	235	119
version 5B	Binary	699	112	52	125
	ADR	699	226	99	138
	Medication	699	163	108	123
	ADR+Medication	699	279	173	113
version 6A	Binary	641	82	110	155
	ADR	641	203	237	161
	Medication	641	139	220	147
	ADR+Medication	641	265	367	127
version 6B	Binary	669	99	82	138
	ADR	669	213	153	151
	Medication	669	151	160	135
	ADR+Medication	669	273	261	119

ADR: adverse drug reaction, TN: true negative, TP: true positive, FN: false fegative, FP: false positive

Chapter 7

Artificial intelligence in cardiovascular imaging: state-of-the-art and implications for the imaging cardiologist

Klaske R. Siegersma, Tim Leiner, Derek P. Chew, Yolande Appelman, Leonard Hofstra, Johan W. Verjans

Netherlands Heart Journal 2019; 27 (9): 403-413

ABSTRACT

Healthcare, conceivably more than any other area of human endeavour, has the greatest potential to be affected by artificial intelligence (AI). This potential has been shown by several reports that demonstrate equal or superhuman performance in medical tasks that aim to improve efficiency, diagnosis and prognosis. This review focuses on the state of the art of AI applications in cardiovascular imaging. It provides an overview of the current applications and studies performed, including the potential value, implications, limitations and future directions of AI in cardiovascular imaging.

It is envisioned that AI will dramatically change the way doctors practise medicine. In the short term, it will assist physicians with easy tasks, such as automating measurements, making predictions based on big data, and putting clinical findings into an evidence-based context. In the long term, AI will not only assist doctors, it has the potential to significantly improve access to health and well-being data for patients and their caretakers. This empowers patients. From a physician's perspective, reliable AI assistance will be available to support clinical decision-making. Although cardiovascular studies implementing AI are increasing in number, the applications have only just started to penetrate contemporary clinical care.

INTRODUCTION

Each year, more and more cardiac imaging investigations are being performed.[1] This is driven by multiple factors, such as increased acceptance of imaging, which over the years has played an incremental role in diagnosis, management and monitoring treatment outcome. In addition, imaging has become more widely available, and imaging equipment has become not only more precise, but also faster and cheaper. The improved quality and interpretability of imaging studies has not only led to increased satisfaction for the patient, but could also lead to increased reassurance of the doctor from a clinical and legal perspective. From an economical perspective, the global increase in healthcare costs is in part related to the increasing number of imaging units present in the hospital and thus the increased number of imaging studies performed.[2] However, the expansion of imaging capabilities and subsequent analyses stretches the limits of productivity of the average imaging specialist. Medical artificial intelligence (AI) is a solution for the standardised evaluation of the increasing number of medical images. Scientific literature has started to demonstrate that smart computers utilising AI can provide guidance and assistance during image acquisition and evaluation. This potentially has a significant influence on the physician's workload.

Why is AI promising for medical imaging?

One definition of AI is 'the science of making machines do things that could be considered intelligent when they were performed by human beings', although 'intelligence' itself could be considered a poorly defined term.[3] AI applications are increasingly used to solve problems in healthcare and medicine, as demonstrated by an increasing number of studies using keywords such as 'artificial intelligence', or the methodological references 'machine learning' (ML) and 'deep learning' (DL).[4] The former refers to the development of models where input variables are predefined, e.g. the use of clinical, stress-testing and imaging variables for the prediction of major adverse cardiac events (MACE).[5] The latter type of learning is based on the intrinsic discovery of important features in a multi-layered model set-up, e.g. using echocardiographic images to classify the view.[6]

AI researchers aim to develop and train self-learning models. These models pursue the identification of sophisticated relationships between a given input and corresponding outcome of multiple samples. As alluded to before, the definition of AI varies between experts, but they all refer to implementation of a distinctly human characteristic in models: exploiting previous experience to increase knowledge on how to perform a task in order to enhance decision-making in the future.[7]

The notion of applying AI to medical imaging is fascinating for multiple reasons. First of all, it is becoming apparent that image datasets harbour considerably more useful data than a human can typically process. Secondly, simple tasks, like drawing contours and

subsequent measurements, can be performed by computers more consistently, without interruption and many times faster than by humans. Although the development of useful ML models will take time, it is postulated that the implementation of AI will enable physicians to start working more efficiently.[8,9]

For medical imaging, AI impacts all steps of the imaging chain (Figure 1). The first step is decision-support for selection of the appropriate diagnostic imaging modality. Currently, healthcare is continuously pushing towards evidence-based decision-making and the use of guidelines. AI-based decision-support tools can aid in the selection of the most appropriate imaging test for individual patients. Furthermore, vendors are currently selling the first commercial products that implement ML during the examination of a patient.[6,10] Following acquisition, AI is implemented in image reconstruction (e.g. using low-dose computed tomography, CT, to obtain an optimal anatomical reconstruction[11]), image interpretation and diagnosis (e.g. computer-aided diagnosis of myocardial infarction, MI, in echocardiography[12]). The final step in the imaging chain is to identify relevant prognostic and predictive information from cardiac imaging (e.g. prediction of adverse outcome in patients with pulmonary hypertension[13]).

The concept of personalised medicine is the combination of specific knowledge about an individual patient's characteristics in order to tailor the predicted prognosis, choose treatment based on anticipated response or susceptibility for a specific disease.[14] Truly personalised medicine for multiple diseases is an important goal for the future of healthcare. For instance, direct application of AI would be very suitable for the evaluation of sex and gender differences, which is an important topic in current cardiovascular research. To accomplish this goal, different sets of data in healthcare must be combined: imaging data, electronic health records, biomarker analysis, genetic data, and others.[15] Although there have been attempts at combining different data sources, with some promising results[5,15,16], current research in AI has not yet reached this level of complexity in healthcare. Most published studies have focused on automated segmentation, post-processing and computer-aided diagnosis. Therefore, this review predominantly focuses on narrow AI

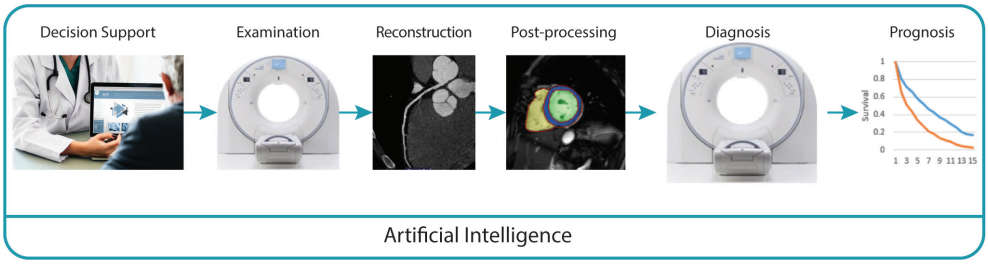


Figure 1 Artificial intelligence is able to impact all steps in the imaging chain.

projects that could prove useful in the near future in cardiovascular imaging. We aim to provide a non-systematic narrative overview of the early applications and studies of the implementation of AI in cardiac imaging, categorised by the different imaging modalities: echocardiography, CT, magnetic resonance imaging (MRI) and nuclear imaging.

IMPLEMENTATION OF AI IN CARDIOVASCULAR IMAGING

Echocardiography

Echocardiography is the most widely used imaging modality in cardiology.[17] The advantages of ultrasound are portability, speed and affordability. However, it is a user-dependent method and intensive training is required in order to achieve accurate interpretation of the acquired data.[18] AI can aid in a more standardised analysis of echocardiographic images, to reduce user dependency. It has already demonstrated the ability to aid in the analysis of echo images, allowing the generation of important cardiac variables on-the-fly with automated classification of echocardiographic views (unpublished; DiA Imaging Analysis/GE Healthcare).

AI has been applied to different steps in the echocardiographic imaging chain. Firstly, during the acquisition of echocardiographic images, automated identification and measurement of the left ventricular wall has been implemented with an ML-based model. Performance of this algorithm is comparable to the traditional 3D echocardiographic methods and cardiac MRI. However, in a minority of pathologies, e.g. congenital disorders or disease with small ventricular cavities, the left ventricular myocardium is not optimally recognised by the implemented algorithm.[19]

Secondly, AI is applied in the post-processing of echocardiographic images. To facilitate a fully automated analysis, algorithmic classification of standard views is essential.[20] Madani et al. showed that a DL model achieves a similar performance in view classification to that of a board-certified echocardiographer (Figure 2).[6] Parameters of cardiac function have been analysed and determined with AI-based models trained with echocardiographic data. Results showed that the determination of left ventricular ejection fraction and longitudinal strain via AI generates similar results to those with expert visual determination.[21] Also, segmentation of the left and right ventricle is of interest, with the goal of automating ejection fraction measurements. The feasibility for the segmentation of the left ventricle using an AI model trained with small training sets has been demonstrated. The accuracy of the segmentation increased when the number of training images used was increased. This result shows an important characteristic of AI; increasing the amount of input data will usually improve the model's performance. However, it should also be noted that using a more diverse dataset of images typically provides more generalisable results.[22] A similar performance was shown for determination of the size and function of the right ventricle. The correlation between automated and convention-

al right ventricular measurements ranged between 0.79 and 0.95 (r-values). However, Bland-Altman analysis showed that both end-diastolic and end-systolic volumes were usually overestimated in automated analyses. Furthermore, this method was semi-automated and required manual tracing of the right ventricular wall in a single frame.[23] All previously mentioned studies applied post-processing steps, performed after data acquisition. Excitingly, on-the-fly echocardiographic analysis has recently been introduced into the software of handheld ultrasound devices, enabling automated analysis of variables during acquisition.

In addition to automated analysis, classifying or diagnosing several cardiac pathologies has been demonstrated. Moghaddasi and Nourian[24] used three different classifiers for the detection of mitral regurgitation. A support vector machine provided the most accurate results for determination of severity (accuracy: >99% for every degree of severity), as evaluated by human interpretation. Automated identification of MI has been enhanced

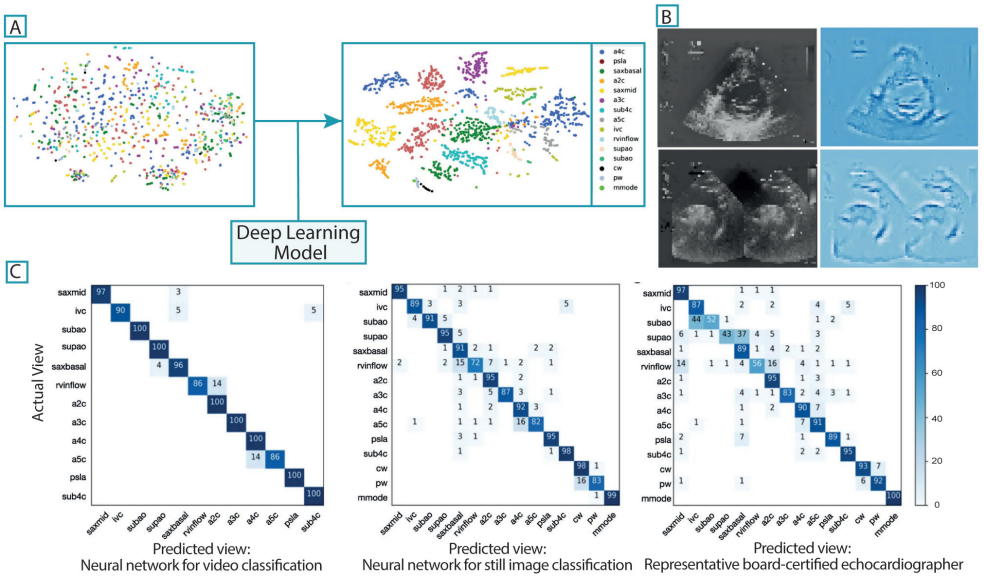


Figure 2 Images obtained from the research performed by Madani et al.[6] A. 2D representation of the differenc echocardiographic views. Different colours represent the different standard echocardiographic views. A deep-learning model enabled classification, which resulted in the clustering as can be seen in the plot on the right. B. The saliency maps (occlusion map not shown). The input pixels weighted most heavily in the neural network's classification of the original images (left). The most important pixels (right) make an outline of relevant structures demonstrating similar patterns that humans use to classify the image. C. The confusion matrices for different classifiers. The actual views are represented on the vertical axes. The horizontal axes represent the classification of views by a neural network with video classification input (c1), a neural network with still images as input (c2) and the classification performed by a board-certified echocardiographer (c3). The numbers in the squares represent the percentage of labels predicted for each category (rounding causes addition to not always add up to 100).

with AI using different input features. Strain rate curves and segmental deformation for identification of MI demonstrated an accuracy of 87%.[25] Another study performed an analysis of MI using texture descriptors derived from the discrete wavelet transforms of the ultrasound signal (accuracy 99.5%).[12] Narula et al.[26] used speckletracking data to discriminate between an athlete's heart and hypertrophic cardiomyopathy with three different classifiers. The models showed increased accuracy when different echocardiographic features were combined compared to single features alone. Although these pathologies are clinically similar, the ML model may present the opportunity to differentiate between phenotypes and modify therapy.[26] Similar echocardiographic data were used to differentiate between patients with restrictive cardiomyopathy and constrictive pericarditis. Although differentiation of these entities using four echocardiographic parameters without AI generates an area under the receiver operating characteristic curve (AUC) of 0.942, an associative memory classifier trained with features from speckle tracking echocardiography in addition to the four echocardiographic features generated an improved AUC of 0.962. While similar results were obtained with and without AI, this study demonstrates that implementation of AI for discrimination of these entities is feasible.[27] Zhang et al.[28] published a study that includes segmentation, calculation of several clinical parameters and diagnosing three different cardiac pathologies in 14,035 echocardiograms. The application of AI was shown to be feasible in many steps along the imaging pathway, e.g. for detection of disease the AUC varied from 0.85 to 0.93.[28]

Another specific diagnostic domain for the implementation of AI is the characterisation of the phenotype of heart failure with preserved ejection fraction (HFpEF). This disease has a heterogeneous profile and its management is limited by the lack of a true gold standard definition.[29] In a study[30] on 100 subjects, both HFpEF patients and healthy, but hypertensive and breathless, control subjects, a classifier had an accuracy of 81% for the classification of patients with HFpEF. This classification was based upon spatial-temporal rest-exercise features, which were partly determined by ML algorithms. This study shows the application of AI in diagnosis and post-processing of imaging, respectively.[30] Shah et al.[31] also used a combination of imaging and clinical variables for classification and prediction of outcome in patients with HFpEF This study showed an AUC between 0.70 and 0.76 during validation. Unsupervised phenomapping of HFpEF patients generated three different phenotypes with a significant difference in endpoints of cardiovascular hospitalisation or death. An unsupervised analysis of a combination of data sources has also been used in the identification of patients with heart failure that benefit from cardiac resynchronisation therapy.[32]

In summary, in the short term AI will likely be implemented in echocardiography for automated segmentation and analysis of left and right ventricle contours and automated calculation of volumetric parameters, thereby reducing the workload of echocardiographic

technicians. Subsequently, classification of disease with AI can be achieved, based solely on echocardiographic images, as well as combining imaging data with clinical variables, supporting clinicians and radiographers on the fly. This will also enable the generation of new hypotheses and lead to better diagnostic and prognostic performance in different cardiovascular pathologies.

Computed tomography

Cardiac CT has made a leap forward in the last decade, focusing on the visualisation of stenosis in the coronary tree, plaque characteristics, coronary calcification and scoring and, more recently, the modelling of flow.[33] Promising opportunities for AI in CT are automated noise reduction, while retaining optimal imaging quality, and the avoidance of invasive coronary angiography (ICA) for determination of significant stenosis.[34,35]

In the context of image acquisition, Wolterink et al.[35] described and validated a method with which to obtain reduced radiation dose CT images by training a DL model. Low-dose CT images were used to estimate the routine-dose CT images. A similar approach with a convolutional neural network was used to determine the calcium score from regular coronary CT angiography (CTA). This obviated the need for calcium score CT and thereby reduced radiation exposure for the patient.[36] Another application of AI in CT is post-processing of the images. Zreik et al.[37] showed that automated segmentation of the left ventricle from coronary CTA with convolutional neural networks is a feasible and reliable option.

A topic that has been extensively studied is the identification of significant coronary stenosis from coronary CTA. Significant coronary stenosis is defined as a fractional flow reserve (FFR) <0.8 determined during ICA. The use of AI replaces the need for invasive measurements and generates clear models of local FFR. Different input features derived from coronary CTA have been used for modelling: physiological features[38], quantitative plaque measurements[39], features calculated from different spatially connected clusters of heart segmentation[40], and geometric features of the coronary anatomy[9,41]. Also features from CT perfusion are being evaluated for use with AI.[42] Currently, non-invasive measurements of FFR are performed with computational fluid dynamics, which is computationally demanding. Substitution of this method by AI was shown to be faster and performance was equally good.[9,38] Improvement of non-invasive determination of FFR was obtained by accounting for partial volume effects with AI. Partial volume effects lead to an overestimation of the vessel lumen area.[43] This development leads to an opportunity to decrease the number of ICAs, while allowing for targeting specific stenosis during ICA and, thus, decreasing the duration of the procedure. Automated identification of coronary artery calcium (CAC) has also been subjected to AI approaches, showing that automated identification of CAC in ECG-gated non-contrast-enhanced CT imaging has

an intra-class correlation coefficient of 0.95. This performance is similar to that of a human expert.[11]

A small number of studies have been performed to determine the diagnostic value of AI in coronary CTA. Zreik et al.[44] obtained an accuracy of 0.77 for the detection and characterisation of coronary plaque. For the detection of stenosis and determination of the anatomical significance, an even higher accuracy of 0.80 was obtained, with a dataset of 163 patients. Kolossvary et al.[45] used a more supervised approach with predefined measurements, so-called radiomics[46] features, to identify coronary plaques with a napkin-ring sign (NRS). This sign is an independent prognostic marker of MACE. A large number of texture features, derived from the radiomics set, are able to differentiate between plaques with and without NRS.[45] Another study used texture features derived from calcium score CT as the input in ML models to discriminate between patients with acute or chronic MI and control subjects. This resulted in an AUC of 0.78 and obviated the need for gadolinium-enhanced MRI.[47] However, it has to be noted that manual segmentation of the coronary plaques and the left ventricular wall was required, creating an extra non-automated action. This limits the possible implementation in the cardiologist's clinical workflow.

Besides automated analysis and diagnostics, prognostic evaluation has been applied in cardiac CT. Survival analysis was performed in different patient groups with a cardiovascular risk. In the classification of all-cause mortality among patients with suspected coronary artery disease (CAD), ML models exhibited a larger AUC (0.79) than the individual clinical and coronary CTA metrics (e.g. Framingham risk score: 0.61, segment stenosis score: 0.64, segment involvement score: 0.64, Duke index: 0.62).[16] A similar prognostic model was developed using coronary CTA features derived from the stenosis. This model generated a risk score for all-cause death and non-fatal MI during a follow-up of >3 years and resulted in an AUC of 0.771. This AUC was higher than for each of the individual conventional coronary CTA variables.[48]

Results of AI-based models are promising for cardiac CT; more specifically there is a great future for coronary CTA, mainly due to the non-invasive nature of CT imaging, which is relatively user-independent and fast. Reducing the radiation dose is a relevant application of AI for patients, but preserving spatial resolution is important to make appropriate diagnostic and possibly prognostic decisions. Another purpose of AI in coronary CTA is reducing the need for ICA by expanding the informational value of the diagnostic images. Given the number of studies that apply AI in CT, this field is expanding and starting to incorporate other data sources in the analysis. This creates valuable models and brings us closer to a world of personalised medicine.

Magnetic resonance imaging

Cardiac MRI is a field that comprises the imaging of many aspects of the heart: anatom-

ical imaging, contractile function, flow imaging, perfusion imaging and, importantly, myocardial characterisation.[49] However, given the many opportunities that cardiac MRI offers with regard to AI applications and the technological methods used in MRI, radiographers that have experience and knowledge of physics and cardiac anatomy are integral to image acquisition and analysis. As a consequence, the quality of cardiac MR images is not only user dependent, but also patient, scanner and vendor dependent.

Automated segmentation of cardiac structures and infarct tissue have been the main topic of interest in cardiac MRI thus far. Several studies have been published on the automated segmentation of cine images[50–54] and automated calculation of cardiac parameters from MRI[55,56]. Multiple software programs are available that perform automated segmentation based upon AI. Algorithms for automated segmentation of enhancement on late gadolinium enhancement imaging were summarised and tested by Karim et al.[57], whose study showed that AI algorithms provided greater accuracy than fixed-model approaches. Beyond performance, the reduction in time is a particularly important characteristic of automated AI-based segmentation[51], as can be seen in Figure 3. Baessler et al.[58] used ML models to select the most important texture features, derived from cine images, to differentiate between patients with MI and control subjects. The use of two texture features in multiple logistic regression generated an AUC of 0.92. Implementation of this model in a clinical setting precludes the need for gadolinium-enhanced cardiac MRI, potentially expanding the eligible patient population and reducing costs. All these studies suggest the feasibility of simplifying further analysis of myocardial tissue in large cardiac MRI datasets. An interesting approach by Snaauw et al.[59] demonstrated to possibility of so-called end-to-end classification of disease on cardiac MR images, without the need for annotation.

Two studies have been reported that perform predictive modelling with cardiac MRI data. First, principal component analysis was used to determine survival in patients with pulmonary hypertension. Input for the analysis was the three-dimensional cardiac motion of the right ventricle. This method showed an AUC of a time-dependent receiver operating characteristic analysis of 0.73 for the inclusion of 3D-MR features in the model, besides clinical, functional and regular MR features and features derived from right sided heart catheterisation (otherwise: AUC 0.60). Median follow-up time was 4.0 years.[13] A second predictive model examined the deterioration of left ventricular function in patients with a repaired tetralogy of Fallot. This study indicated that ML models can be useful for planning early intervention in patients at high risk (AUC: 0.87 for major deterioration). Follow-up duration had a median of 2.7 years.[60]

To conclude, due to major disadvantages that compromise inter- and intra-patient comparability in MRI, the application of diagnostic and prognostic Al in MRI is more challenging than in other imaging modalities. In the short term, the use of Al in cardiac MRI will

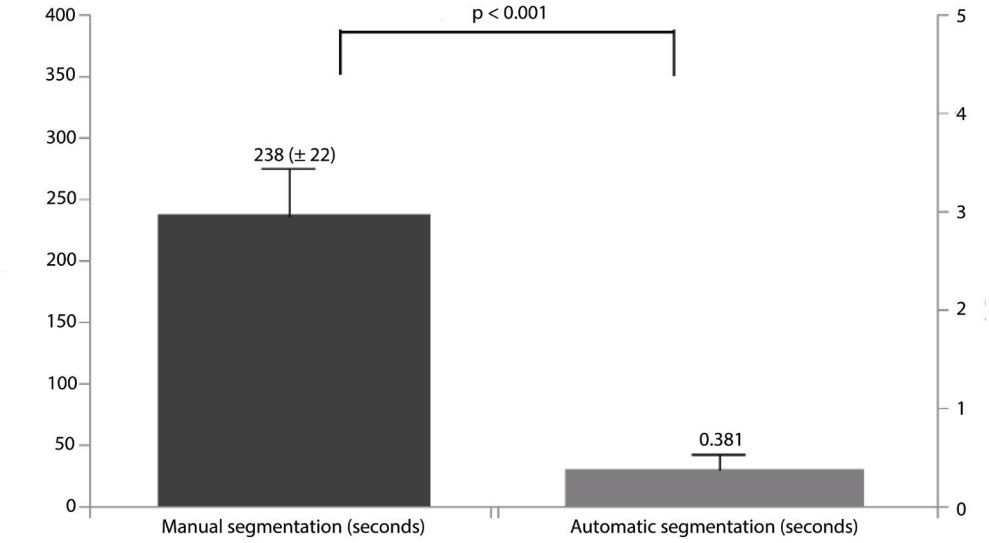


Figure 3 Comparison of processing times of segmentation of the aortic valve in cardiovascular magnetic resonance phase contrast imaging. Automated segmentation used a neural network approach, trained with 150 segmentations. Validation was done in a cohort of 190 segmentations. Automated segmentation times were obtained with GPU acceleration. However, also without GPU acceleration, the average segmentation time was 19.04s. (Images obtained from Bratt et al.[51])

therefore be primarily focusing on automated segmentation and calculation of variables. There are vendors who have incorporated this into their software. Nonetheless, efforts to increase the use of AI for diagnostics and prognostics in CMR continue, with the key challenge of overcoming the between-study comparability of MR images. Methods are being developed and currently further optimised and standardised.

Nuclear imaging

Nuclear imaging of the heart is used to assess perfusion defects within the myocardial wall. Myocardial perfusion single-photon emission computed tomography (SPECT) and positron emission tomography (PET) are methods for cardiac nuclear imaging, although the latter is rarely used in clinical practice due to its costliness, among other reasons. Whereas PET is based on the simultaneous detection of two opposite annihilation photons, SPECT uses gamma rays emitted by a radioactive tracer to reconstruct tissue with uptake. Both nuclear imaging methods can be combined with MRI or CT, which has shown to improve their clinical value, although cardiac PET-MRI has only just started to be used in larger centres.[61,62]

The automated analysis of SPECT is a growing field of interest for research. Normal and abnormal myocardium in CAD can be classified with AI-based models, with performance reported to be similar to the visual analysis of SPECT images.[63] Also, the detection of lo-

cations with abnormal myocardium has been investigated. An artificial neural network, trained with expert interpretations of SPECT images, improved the identification of stress (AUC: 0.92), rest defects (AUC: 0.97) and stress-induced ischaemia (AUC: 0.97) compared to conventional scoring; the AUC of the summed stress score, summed difference score and the summed rest score was 0.82, 0.75 and 0.91, respectively.[64] Another study by this group compared an improved version of the neural network to the older version, showing that retraining of the model improved the identification of ischaemia. The AUC increased to 0.96.[65]

The accuracy of SPECT can be boosted by the integration of clinical data and quantitative imaging features in an ML model. The diagnostic accuracy in the detection of obstructive CAD was improved with an ML model with quantitative and clinical features. This model generated a marginally better result than a model with solely clinical features (accuracy: 79.4% vs 75.7%). The performance of the model was similar to the visual analysis of one experienced reader (78.5%) and better than another (73.5%).[66] Also, Betancur et al.[8] examined the automated prediction of obstructive CAD. DL models were trained with the raw and quantitative perfusion polar maps. AI-based models showed a higher AUC (0.80) for prediction of CAD than the current clinical method (0.78) in 1638 subjects. Another study showed utility in aiding decision-making for cardiac interventions. SPECT data, merged with functional and clinical features, were used to predict the necessity for revascularisation. The results of this study showed that an ML approach (AUC: 0.81) was comparable to or better than two experienced readers (AUC: 0.81 and 0.72) in the prediction of the need for revascularisation.[67]

MACE were also studied in 2619 patients who were referred for myocardial perfusion imaging. This risk analysis was based upon an ML model that combined clinical information with myocardial perfusion SPECT data. This model showed a higher AUC than a model with solely imaging features (AUC: 0.81 vs 0.78).[5] Another study by Haro Alonso et al.[68] compared an ML model with baseline logistic regression for the prediction of cardiac death. Patients were selected if they had undergone myocardial perfusion SPECT and imaging parameters were used for modelling. The study showed that baseline logistic regression (AUC: 0.77) was outperformed by all ML models, with the support vector machine generating the highest AUC (0.83). ML models have also been used in cardiac PET. However, in this case PET variables were used as the output classification, using demographic, clinical and functional variables as input. ML models were superior to logistic regression for the identification of myocardial ischaemia, based upon PET images, and selection of patients at risk for MACE.[69]

To conclude, short-term applications for AI in nuclear imaging are predominantly focused on automated detection of perfusion defects in the myocardial wall. Because nuclear imaging can be easily combined with CT or MRI, this enables enhanced fusion of multiple

data sources in addition to clinical data. Such methods have been shown to improve the performance of diagnostic and predictive models. However, in the long term, the high radiation exposure during nuclear imaging remains a limitation in the cardiac clinic, and hence also to the penetration of AI into this imaging modality.

CONCLUSIONS AND FUTURE DIRECTIONS

Cardiovascular imaging has shown remarkable advancements in the last few decades, leading to detailed imaging of not only structural, but also physiological and even molecular characteristics of the cardiovascular system. The advancement of AI creates opportunities in healthcare to obtain more sophisticated information from imaging, and to find patterns in available data sources that are too complex for the human brain. Intuitively, it is more a question of when, rather than if, AI technology will offer significant help for the cardiologist, and in particular the imaging specialist. Applications are already being implemented in the clinical workflow based on research that shows equal or better performance than analysis by the physician or conventional (semi-)automated methods, with the aim of reducing the workload of the physician and enhancing decision-making. The use of AI in cardiac CT applies to many steps of the imaging chain, whereas the application of AI in cardiac MRI has so far primarily focused on automated segmentation of anatomical structures of the heart. In addition, nuclear imaging and echocardiography have used predictive and prognostic modelling. Nevertheless, the implementation of AI faces significant challenges. There are efforts underway to improve the comparability of imaging modes. Automated segmentation or extraction of imaging features is likely to be 'solved' first, and this will help to standardise and accelerate the analysis of large datasets.

Despite promising results, the implementation of AI in contemporary cardiovascular healthcare has been limited to date.[70,71] Several reasons contribute to this observation. First, regulatory bodies, like the American Food and Drug Administration, have difficulty with the regulation and approval of software based on AI.[72] This delays the allotting of certification marks and the introduction of products on the consumer market. Second, the added value of AI in clinical care remains to be determined and established. No studies have been performed that show that the implementation of AI indeed leads to higher quality of care, lower healthcare costs or improved patient outcomes.[15] Furthermore, due to the 'black box' used in many ML models and the dependency on input data for the performance of a model, it is difficult to replicate or explain experiments. Repeating studies and validating designed ML models will be important before routine implementation in clinical practice.[73] This requires sharing of data and model settings, which is uncommon in clinical research; also patient privacy and compliance with regulations regarding patient data are critical considerations. Third, physicians are not yet prepared

for the implementation of AI in the daily clinical setting. Trust in these new technologies has to be built, supported by efforts towards transparency and explainability.[74] Fourth, the datasets used in the described studies are commonly relatively small. A large range of different patients must be included in studies to develop appropriate models. Diversity in ethnicity, gender and age must be guaranteed to build widely applicable models. This includes the data used during training, validation and testing of the model. Furthermore, a standardised method for storing or extraction of information in the electronic health record should be developed. The use of free text should be reduced to enhance data analysis and application of AI.

There is also an opportunity for future research to focus on the implementation of data from multiple sources in ML models, including biomarkers, genomics, proteomics and metabolomics.[15] This can improve predictive value of ML models and create personalised healthcare for patients. Text mining and improvement of the predictive value of free text analysis are being explored[75,76], but standardised reporting can clearly facilitate the implementation of AI worldwide. The implementation of multiple sources in ML models can also contribute in deciding whether to refer a patient for cardiac imaging, e.g. immediate therapeutic decision-making based on CT data without the need for ICA[41].

Current AI technology is considered to be 'narrow', meaning it is good at one particular task, and it is only as good as the dataset that trained it. AI has made remarkable progress, and despite a clear peak of potentially inflated expectations, the number of translational studies that implement so-called narrow AI is slowly growing, with some results already showing performance that is equal to or better than that of conventional methods (e.g. Motwani et al.[16]) or expert analysis (e.g. Arsanjani et al.[66]). It will take more than a few decades before we are able to achieve so-called general, human-like AI. It will undoubtedly take time for the adoption of such methods in daily clinical practice, where decisions are complex. Moreover, practice is relatively conservative in the face of ethical and medico-legal considerations.[74]

Physicians need to realise that AI is a tool that will not replace many tasks in the short term, but will likely enhance diagnostic and decision-making capacity. Human performance will be augmented, and it is likely this will improve the outcome of patients through better diagnosis, fewer errors and significant time-saving that could help us create more productive patient-doctor interactions.

REFERENCES

1. RIVM. Trend in aantallen verrichting. Diagnostiek. Published 2018. Accessed October 25, 2021. https://www.rivm.nl/medische-stralingstoepassingen/trends-en-stand-van-zaken/diagnostiek#Trend in aantallen verrichtingen
2. Papanicolas I, Woskie LR, Jha AK. Health Care Spending in the United States and Other High-Income Countries. *JAMA*. 2019;319(10):1024-1039. doi:10.1001/jama.2018.1150
3. Minsky M. Why People Think Computers Can't. *AI Mag*. 1982;3(4):3-15.
4. Corlan AD. Medline trend: automated yearly statistics of PubMed results for any query. Published 2004. Accessed November 12, 2018. http://dan.corlan.net/medline-trend.html
5. Betancur J, Otaki Y, Motwani M, et al. Prognostic Value of Combined Clinical and Myocardial Perfusion Imaging Data Using Machine Learning. *JACC Cardiovasc Imaging*. 2018;11(7):1000-1009. doi:10.1016/j.jcmg.2017.07.024
6. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *npj Digit Med*. 2018;1(1):1-8. doi:10.1038/s41746-017-0013-1
7. Russell S, Norvig P, eds. Introduction. In: *Artificial Intelligence: A Modern Approach*. 3rd ed. Malaysia: Pearson Education Limited; 2016:1-30.
8. Betancur J, Commandeur F, Motlagh M, et al. Deep Learning for Prediction of Obstructive Disease From Fast Myocardial Perfusion SPECT: A Multicenter Study. *JACC Cardiovasc Imaging*. 2018;11(11):1654-1663. doi:10.1016/j.jcmg.2018.01.020
9. Coenen A, Kim YH, Kruk M, et al. Diagnostic accuracy of a machine-learning approach to coronary computed tomographic angiography–Based fractional flow reserve result from the MACHINE Consortium. *Circ Cardiovasc Imaging*. 2018;11(6):1-11. doi:10.1161/CIRCIMAGING.117.007217
10. Graff CG, Sidky EY. Compressive sensing in medical imaging. *Appl Opt*. 2015;54(8):C23-C44.
11. Wolterink JM, Leiner T, Takx RAP, Viergever MA, Išgum I. Automatic Coronary Calcium Scoring in Non-Contrast-Enhanced ECG-Triggered Cardiac CT With Ambiguity Detection. *IEEE Trans Med Imaging*. 2015;34(9):1867-1878. doi:10.1109/TMI.2015.2412651
12. Sudarshan VK, Ng EYK, Acharya UR, Chou SM, Tan RS, Ghista DN. Computer-aided diagnosis of Myocardial Infarction using ultrasound images with DWT, GLCM and HOS methods: A comparative study. *Comput Biol Med*. 2015;62(2015):86-93. doi:10.1016/j.compbiomed.2015.03.033
13. Dawes TJW, De Marvao A, Shi W, et al. Machine learning of threedimensional right ventricular motion enables outcome prediction in pulmonary hypertension: A cardiac MR imaging study. *Radiology*. 2017;283(2):381-390. doi:10.1148/radiol.2016161315
14. Redekop WK, Mladsi D. The Faces of Personalized Medicine: A Framework for Understanding Its Meaning and Scope. *Value Heal*. 2013;16(6 SUPPL.):S4-S9. doi:10.1016/j.jval.2013.06.005
15. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: Promise and challenges. *Nat Rev Cardiol*. 2016;13(6):350-359. doi:10.1038/nrcardio.2016.42
16. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017;38:500-507. doi:10.1093/eurheartj/ehw188
17. Hasselberg NE, Edvardsen T. Ultrasound/

echocardiography. In: Nieman K, Gaemperli O, Lancellotti P, Plein S, eds. *Advanced Cardiac Imaging*. 1st ed. Woodhead Publishing; 2015:15-46.
18. Feigenbaum H. Evolution of echocardiography. *Circulation*. 1996;93(7):1321-1327. doi:10.1161/01.CIR.93.7.1321
19. Tamborini G, Piazzese C, Lang RM, et al. Feasibility and Accuracy of Automated Software for Transthoracic Three-Dimensional Left Ventricular Volume and Function Analysis: Comparisons with Two-Dimensional Echocardiography, Three-Dimensional Transthoracic Manual Method, and Cardiac Magnetic Resona. *J Am Soc Echocardiogr*. 2017;30(11):1049-1058. doi:10.1016/j.echo.2017.06.026
20. Khamis H, Zurakhov G, Azar V, Raz A, Friedman Z, Adam D. Automatic apical view classification of echocardiograms using a discriminative learning dictionary. *Med Image Anal*. 2017;36:15-21. doi:10.1016/j.media.2016.10.007
21. Knackstedt C, Bekkers SCAM, Schummers G, et al. Fully Automated Versus Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain the FAST-EFs Multicenter Study. *J Am Coll Cardiol*. 2015;66(13):1456-1466. doi:10.1016/j.jacc.2015.07.052
22. Carneiro G, Nascimento JC. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(11):2592-2607. doi:10.1109/TPAMI.2013.96
23. Medvedofsky D, Addetia K, Hamilton J, Leon Jimenez J, Lang RM, Mor-Avi V. Semi-automated echocardiographic quantification of right ventricular size and function. *Int J Cardiovasc Imaging*. 2015;31(6):1149-1157. doi:10.1007/s10554-015-0672-4
24. Moghaddasi H, Nourian S. Automatic assessment of mitral regurgitation severity based on extensive textural features on 2D echocardiography videos. *Comput Biol Med*. 2016;73:47-55. doi:10.1016/j.compbiomed.2016.03.026
25. Tabassian M, Alessandrini M, Herbots L, et al. Machine learning of the spatio-temporal characteristics of echocardiographic deformation curves for infarct classification. *Int J Cardiovasc Imaging*. 2017;33(8):1159-1167. doi:10.1007/s10554-017-1108-0
26. Narula S, Shameer K, Salem Omar AM, Dudley JT, Sengupta PP. Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. *J Am Coll Cardiol*. 2016;68(21):2287-2295. doi:10.1016/j.jacc.2016.08.062
27. Sengupta PP, Huang YM, Bansal M, et al. Cognitive Machine-Learning Algorithm for Cardiac Imaging; A Pilot Study for Differentiating Constrictive Pericarditis from Restrictive Cardiomyopathy. *Circ Cardiovasc Imaging*. 2016;9(6):1-10. doi:10.1161/CIRCIMAGING.115.004330
28. Zhang J, Gajjala S, Agrawal P, et al. Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*. 2018;138(16):1623-1635. doi:10.1161/CIRCULATIONAHA.118.034338
29. The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J*. 2016;37:2129-2200. doi:10.1093/eurheartj/ehw128
30. Tabassian M, Sunderji I, Erdei T, et al. Diagnosis of Heart Failure With Preserved Ejection Fraction: Machine Learning of Spatiotemporal Variations in Left Ventricular Deformation. *J Am Soc Echocardiogr*. 2018;31(12):1272-1284.e9. doi:10.1016/j.echo.2018.07.013
31. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for novel classification of heart failure with preserved ejection

fraction. *Circulation*. 2015;131(3):269-279. doi:10.1161/CIRCULATIONAHA.114.010637
32. Cikes M, Sanchez-Martinez S, Claggett B, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *Eur J Heart Fail*. 2019;21(1):74-85. doi:10.1002/ejhf.1333
33. Nieman K, Coenen A, Dijkshoorn ML. Computed Tomography. In: Nieman K, Gaemperli O, Lancellotti P, Plein S, eds. *Advanced Cardiac Imaging*. 1st ed. Woodhead Publishing; 2015:97-125.
34. Budoff MJ, Dowe D, Jollis JG, et al. Diagnostic performance of 64-multidetector row coronary computed tomographic angiography for evaluation of coronary artery stenosis in individuals without known coronary artery disease: results from the prospective multicenter ACCURACY (Assessment by Coronary Computed Tomographic Angiography of Individuals Undergoing Invasive Coronary Angiography) trial. *J Am Coll Cardiol*. 2008;52(21):1724-1732. doi:10.1016/j.jacc.2008.07.031
35. Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative adversarial networks for noise reduction in low-dose CT. *IEEE Trans Med Imaging*. 2017;36(12):2536-2545. doi:10.1109/TMI.2017.2708987
36. Wolterink JM, Leiner T, de Vos B, van Hamersvelt RW, Viergever MA, Isgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med Image Anal*. 2016;34:123-136. doi:http://dx.doi.org/10.1016/j.media.2016.04.004
37. Zreik M, Leiner T, De Vos BD, Van Hamersvelt RW, Viergever MA, Isgum I. Automatic segmentation of the left ventricle in cardiac CT angiography using convolutional neural networks. In: *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. ; 2016:40-43.
38. Itu L, Rapaka S, Passerini T, et al. A machine-learning approach for computation of fractional flow reserve from coronary computed tomography. *J Appl Physiol*. 2016;121(1):42-52. doi:10.1152/japplphysiol.00752.2015
39. Dey D, Gaur S, Ovrehus KA, et al. Integrated prediction of lesion-specific ischaemia from quantitative coronary CT angiography using machine learning: A multicentre study. *Eur Radiol*. 2018;28(6):2655-2664. doi:10.1007/s00330-017-5223-z
40. Zreik M, Lessmann N, van Hamersvelt RW, et al. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. *Med Image Anal*. 2018;44:72-85. doi:10.1016/j.media.2017.11.008
41. Tesche C, Vliegenthart R, Duguay TM, et al. Coronary Computed Tomographic Angiography-Derived Fractional Flow Reserve for Therapeutic Decision Making. *Am J Cardiol*. 2017;120(12):2121-2127. doi:10.1016/j.amjcard.2017.08.034
42. Han D, Lee JH, Rizvi A, et al. Incremental role of resting myocardial computed tomography perfusion for predicting physiologically significant coronary artery disease: A machine learning approach. *J Nucl Cardiol*. 2018;25(1):223-233. doi:10.1007/s12350-017-0834-y
43. Freiman M, Nickisch H, Prevrhal S, et al. Improving CCTA-based lesions' hemodynamic significance assessment by accounting for partial volume modeling in automatic coronary lumen segmentation. *Med Phys*. 2017;44(3):1040-1049. doi:10.1002/mp.12121
44. Zreik M, Hamersvelt RW Van, Wolterink JM, Leiner T, Viergever MA, Isgum I. A Recurrent CNN for Automatic Detection and Classification of Coronary Artery Plaque and Stenosis in Coronary CT Angiography. *IEEE Trans Med Imaging*. Published online 2019:1-11. doi:10.1109/TMI.2018.2883807
45. Kolossváry M, Karády J, Szilveszter B, et al. Radiomic Features Are Superior to Conventional Quantitative Computed Tomographic Metrics to Identify Coronary

Plaques with Napkin-Ring Sign. *Circ Cardiovasc Imaging*. 2017;10(12). doi:10.1161/CIRCIMAGING.117.006843
46. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5(4006). doi:10.1038/ncomms5006
47. Mannil M, Von Spiczak J, Manka R, Alkadhi H. Texture Analysis and Machine Learning for Detecting Myocardial Infarction in Noncontrast Low-Dose Computed Tomography: Unveiling the Invisible. *Invest Radiol*. 2018;53(6):338-343. doi:10.1097/RLI.0000000000000448
48. van Rosendael AR, Maliakal G, Kolli KK, et al. Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry. *J Cardiovasc Comput Tomogr*. 2018;12(3):204-209. doi:10.1016/j.jcct.2018.04.011
49. Ferreira VM, Robson MD, Karamitsos TD, Bissell MM, Tyler DJ, Neubauer S. Magnetic resonance imaging. In: Nieman K, Gaemperli O, Lancellotti P, Plein S, eds. *Advanced Cardiac Imaging*. 1st ed. Woodhead Publishing; 2015:127-169.
50. Avendi MR, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med Image Anal*. 2016;30:108-119. doi:10.1016/j.media.2016.01.005
51. Bratt A, Kim J, Pollie M, et al. Machine learning derived segmentation of phase velocity encoded cardiovascular magnetic resonance for fully automated aortic flow quantification. *J Cardiovasc Magn Reson*. 2019;0:1-11. doi:10.1186/s12968-018-0509-0
52. Ngo TA, Lu Z, Carneiro G. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. *Med Image Anal*. 2017;35:159-171. doi:10.1016/j.media.2016.05.009
53. Tan LK, Liew YM, Lim E, McLaughlin RA. Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine MR sequences. *Med Image Anal*. 2017;39:78-86. doi:10.1016/j.media.2017.04.002
54. Zheng Q, Delingette H, Duchateau N, Ayache N. 3-D Consistent and Robust Segmentation of Cardiac Images by Deep Learning With Spatial Propagation. *IEEE Trans Med Imaging*. 2018;37(9):2137-2148. doi:10.1109/TMI.2018.2820742
55. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing. *J Cardiovasc Magn Reson*. 2018;20(1):1-12. doi:10.1186/s12968-018-0471-x
56. Suinesiaputra A, Sanghvi MM, Aung N, et al. Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. *Int J Cardiovasc Imaging*. 2018;34(2):281-291. doi:10.1007/s10554-017-1225-9
57. Karim R, Bhagirath P, Claus P, et al. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images. *Med Image Anal*. 2016;30:95-107. doi:10.1016/j.media.2016.01.004
58. Baessler B, Mannil M, Oebel S, Maintz D, Alkadhi H, Manka R. Subacute and chronic left ventricular myocardial scar: Accuracy of texture analysis on nonenhanced cine MR images. *Radiology*. 2018;286(1):103-112. doi:10.1148/radiol.2017170213
59. Snaauw G, Gong D, Maicas G, et al. End-To-End Diagnosis And Segmentation Learning From Cardiac Magnetic Resonance Imaging. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. ; 2019:802-805. doi:10.1109/ISBI.2019.8759276

60. Samad MD, Wehner GJ, Arbabshirani MR, et al. Predicting deterioration of ventricular function in patients with repaired tetralogy of Fallot using machine learning. *Eur Heart J Cardiovasc Imaging*. 2018;19(7):730-738. doi:10.1093/ehjci/jey003
61. Knaapen P, Lubberink M. Positron emission tomography. In: Nieman K, Gaemperli O, Lancellotti P, Plein S, eds. *Advanced Cardiac Imaging*. 1st ed. Woodhead Publishing; 2015:71-95.
62. Buechel RR, Kaufmann PA, Gaemperli O. Single-Photon Emission Computed Tomography. In: Nieman K, Gaemperli O, Lancellotti P, Plein S, eds. *Advanced Cardiac Imaging*. 1st ed. Woodhead Publishing; 2015:47-69.
63. Driessen RS, Raijmakers PG, Danad I, et al. Automated SPECT analysis compared with expert visual scoring for the detection of FFR-defined coronary artery disease. *Eur J Nucl Med Mol Imaging*. 2018;45(7):1091-1100. doi:10.1007/s00259-018-3951-1
64. Nakajima K, Kudo T, Nakata T, et al. Diagnostic accuracy of an artificial neural network compared with statistical quantitation of myocardial perfusion images: A Japanese multicenter study. *Eur J Nucl Med Mol Imaging*. 2017;44(13):2280-2289. doi:10.1007/s00259-017-3834-x
65. Nakajima K, Okuda K, Watanabe S, et al. Artificial neural network retrained to detect myocardial ischemia using a Japanese multicenter database. *Ann Nucl Med*. 2018;32(5):303-310. doi:10.1007/s12149-018-1247-y
66. Arsanjani R, Xu Y, Dey D, et al. Improved accuracy of myocardial perfusion SPECT for detection of coronary artery disease by machine learning in a large population. *J Nucl Cardiol*. 2013;20(4):553-562. doi:10.1007/s12350-013-9706-2
67. Arsanjani R, Dey D, Khachatryan T, et al. Prediction of revascularization after myocardial perfusion SPECT by machine learning in a large population. *J Nucl Cardiol*. 2015;22(5):877-884. doi:10.1007/s12350-014-0027-x
68. Haro Alonso D, Wernick MN, Yang Y, Germano G, Berman DS, Slomka P. Prediction of cardiac death after adenosine myocardial perfusion SPECT based on machine learning. *J Nucl Cardiol*. 2019;26(5):1746-1754. doi:10.1007/s12350-018-1250-7
69. Juarez-Orozco LE, Knol RJJ, Sanchez-Catasus CA, Martinez-Manzanera O, van der Zant FM, Knuuti J. Machine learning in the integration of simple variables for identifying patients with myocardial ischemia. *J Nucl Cardiol*. 2020;27(1):147-155. doi:10.1007/s12350-018-1304-x
70. Lawrence C, Pencina MJ. On Deep Learning for Medical Image Analysis. *J Am Med Assoc*. 2018;320(11):1101-1102. doi:10.1001/jama.2018.11100
71. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am J Med*. 2018;131(2):129-133. doi:10.1016/j.amjmed.2017.10.035
72. U.S. Food & Drug Administration. Digital Health Innovation Action Plan. Published online 2017:1-7.
73. Editorials. AI diagnostics need attention. *Nature*. 2018;555:285-286.
74. Krittanawong C, Zhang HJ, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. *J Am Coll Cardiol*. 2017;69(21):2657-2664. doi:10.1016/j.jacc.2017.03.571
75. Mortazavi BJ, Desai N, Zhang J, et al. Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures. *IEEE J Biomed Heal Informatics*. 2017;21(6):1719-1729. doi:10.1109/JBHI.2017.2675340
76. Wagholikar KB, Fischer CM, Goodson A, et al. Extraction of Ejection Fraction from Echocardiography Notes for Constructing a Cohort of Patients having Heart Failure with reduced Ejection Fraction (HFrEF). *J Med Syst*. 2018;42(11). doi:10.1007/s10916-018-1066-7

Chapter 8

Improving the classification of women at high risk of coronary artery disease with logistic regression and gradient boosting using a regular care database

Klaske R. Siegersma, Leonard Hofstra, Yolande Appelman, Jan-Walter Benjamins, Pim van der Harst, Igor I. Tulevski, Hester M. den Ruijter, G. Aernout Somsen[*], *N. Charlotte Onland-Moret*[*]

[*] These authors contributed equally

In preparation

ABSTRACT

Background The pre-test probability (PTP) risk stratification tool is currently used to guide the decision to refer patients for (non-) invasive additional diagnostic testing for obstructive coronary artery disease (CAD). Based on this probability patients are either referred or not. The PTP slightly underestimates the presence of CAD, and sex-specific performance metrics have not been published. Additional cardiovascular risk factors are used to improve decision-making in patients with CAD. However, these risk factors are not structurally assessed. The wealth of additional information currently available in electronic health records may add to the performance of the PTP.

Aim To show sex-specific performance and calibration of the PTP for CAD and demonstrate whether performance improves with the addition of variables and use of machine learning models.

Methods The study population consisted of 34,524 patients (19,141 women, 55.4%) with chest pain or dyspnoea, that visited a diagnostic outpatient cardiology clinic for the first time. Data extraction from electronic health records and feature engineering resulted in 521 features. Outcome classification was a diagnosis of CAD during or within six weeks after the first consult made by the cardiologist. PTP of CAD was classified according to the 2019 ESC guidelines for chronic coronary syndromes. A gradient boosting and Lasso logistic regression model were fitted on 75% of the data and performance was evaluated on the remaining 25%. Performance of the models was compared to the pre-test probability with the net reclassification index.

Results Mean age of the study population was 55 years. In total, 2719 (7.9%, 1253 women, 46%) patients had a diagnosis of CAD. In women with CAD, the performance of the PTP was poor (AUC=0.64), and 21.9% of women with a diagnosis of CAD were classified in the lowest category (<%5). Classification performance improved for both sexes with the Lasso logistic regression and additional features (AUC Lasso logistic regression: 0.78, 95% CI 0.77-0.79, gradient boosting: 0.78, 95% CI 0.77-0.79 versus AUC PTP 0.65, 95% CI 0.63-0.66 with a 0.05 cut-off). Compared to the PTP, women benefited most from addition of diagnostic variables, in which Lasso logistic regression performed better than a gradient boosting model (NRI Lasso logistic regression in women and men; 0.44 vs 0.35 respectively).

Discussion The use of a Lasso logistic regression with additional diagnostic variables improved classification of CAD in patients with chest pain and dyspnoea compared to the PTP, especially in women, in which the PTP tends to underestimate the risk of CAD.

INTRODUCTION

The burden of heart disease has been increasing worldwide, primarily because of unhealthy lifestyles and population ageing.[1,2] Consequently, healthcare costs are rising. Reducing healthcare costs can be achieved by a smarter use of resources. Accurate patient profiling can aid in matching the right diagnostics, therapies or interventions to the right patient, and thus reducing unnecessary diagnostic procedures and ineffective therapy or interventions[3]. Hence, better patient profiling can lead to a reduction in healthcare costs. The latest European Society of Cardiology (ESC, 2019) guidelines for chronic coronary syndromes[4] assign a prominent role to the pre-test probability (PTP) for obstructive coronary artery disease (CAD) in profiling and risk stratification of patients with symptoms of chest pain or dyspnoea. This probability is used to guide the referral for non-invasive and invasive diagnostic testing for obstructive CAD. The PTP is based on sex, age and type of chest pain or the presence of dyspnoea[5]. It has been validated in several studies[6,7] and showed to be well calibrated for a combined endpoint of obstructive CAD based on diagnostic information from coronary computed tomographic angiography (CCTA) or invasive coronary angiography (CAG).[7] However, the PTP slightly underestimated the probability of CAD, based on information of CCTA alone, specifically in women.[6,7] On top, no performance metrics of the PTP have been published in both sexes separately. This limits the interpretation of sex bias. Hence, reporting of sex-stratified results is still scarce, despite the need for a sex-specific view on cardiovascular disease.[8,9]

No studies to date have combined the patient information from a diagnostic intake in a prediction model for the presence of CAD, although the need for improvement of the PTP has been acknowledged through evaluation of additional risk factors.[4,7] In the ESC guidelines, the incorporation of risk factors was done through additional evaluation of risk modifiers, in patients with an intermediate PTP of CAD (PTP: 5-15%)[4,7,10], e.g. ultrasound features of atherosclerosis outside of the coronaries[11] and a positive ECG stress test[12,13]. No sex-specific risk factors have been added to this evaluation, despite knowledge on their influence on CAD.[14,15]

The digitalized structure of healthcare nowadays enables the mining of enormous amounts of regular care data on men and women. These data can aid in sex-specific evaluation of the PTP in a regular care patient population.[16] In addition, the rapidly evolving field of artificial intelligence (AI) has led to the development of accurate models for patient profiling based on multidimensional data.[17,18] The combination of both techniques enhances new discoveries in cardiovascular disease mechanisms and identification of sex differences. It has also produced improved risk-prediction models for mortality in patients with acute coronary syndromes[19] or CAD[20], and survival after echocardiography in a large regional database[21].

Therefore, the aims of these study are to study the performance of the PTP[4] in a real-life

setting in women and men separately. In addition, we aim to evaluate whether the PTP improves by addition of other diagnostic variables in an AI model as well as by the use of sex-specific models. This will be studied in a regular care database from a cardiology outpatient clinic.[22]

METHODS

Patient selection

The Cardiology Centers of the Netherlands (CCN) database includes 109,151 individual patients and has been described in detail previously.[22] In brief, each patient referred to one of the CCN clinics underwent a standardized diagnostic workup, which consisted of transthoracic echocardiography (TTE) and ultrasound imaging of the carotid arteries, electrocardiography (ECG) at rest and during exercise and a basic laboratory test. Furthermore, each patient has a consult with a specialized nurse, that encompasses self-reported anthropometrics, symptoms, cardiovascular and sex-specific risk factors, comorbidities, previous diagnoses, medical and cardiovascular history and medication use.[22] All patients that entered the clinic with chest pain or dyspnoea for the first were included in the selected patient population. Patients were included in the dataset when the intake diagnostics and consult took place on the same day. Patients were excluded if they had with a history of cardiovascular disease or medical risk factors for coronary heart disease, i.e. cardiomyopathy, atherosclerosis, peripheral vascular disease. Figure 1 shows the process of patient selection. All available variables from the diagnostic intake in the CCN's EHR were included in the dataset (Supplementary table 1).

Outcome classification

CCN uses the nationwide system of the Diagnosis Treatment Combination (DTC, in Dutch: DOT, Diagnose-Behandel Combinatie op weg naar transparantie) to fund care delivered by hospitals and medical specialists. Patient outcome was classified according to the presence of a diagnosis of CAD as determined and registered in the EHR by the cardiologist. Therefore, the selected DTCs were stable angina pectoris, instable angina pectoris, myocardial infarction and non-ST elevated myocardial infarction. Subjects with one of these diagnoses during or within six weeks after their intake were included in the dataset as patients with CAD. All others were included as patients without CAD.

Feature engineering and cleaning of EHR data

All patient variables were derived from the EHR. Medication use, previous diagnoses at the baseline consult and categorical features were dummy encoded. Free text fields for description of chest pain, for the conclusion and reason to stop of the stress ECG, and for indication and conclusion of the rest ECG were encoded using string matching with

specific search terms and presence or absence of negation. An example for chest pain and for the reason to end the stress ECG is included in Supplementary figure 1. Based on the filed chest pain characteristics, a classification of chest pain (non-anginal, atypical, typical[23]) was appointed to each patient. Remaining free text variables were removed from the dataset (n = 59). Measured lab values were included in two ways; as the original measured value and as a dummy variable demonstrating if the lab value was within range of reference values.[24] PTP was determined for each individual based on type of chest pain or presence of dyspnoea, age and sex[5] and was turned from a percentage into a probability by dividing by 100 to enhance comparison with outcomes of the other models. After handling of missing values, all categorical features were turned into dummy variables and features with only one level in either men or women were removed (n = 40). This resulted in a final feature set of 520 features. An overview of the feature classes is shown in Supplementary table 1. Feature engineering was done in the R programming language (R Foundation for Statistical Computing, https://ww.R-project.org , version 4.0.2), RStudio (RStudio: Integrated Development Environment for R, http://www.rstudio.com/, version 1.3.1093) and Python (Python Software Foundation, https://www.python.org, version 3.7.9).

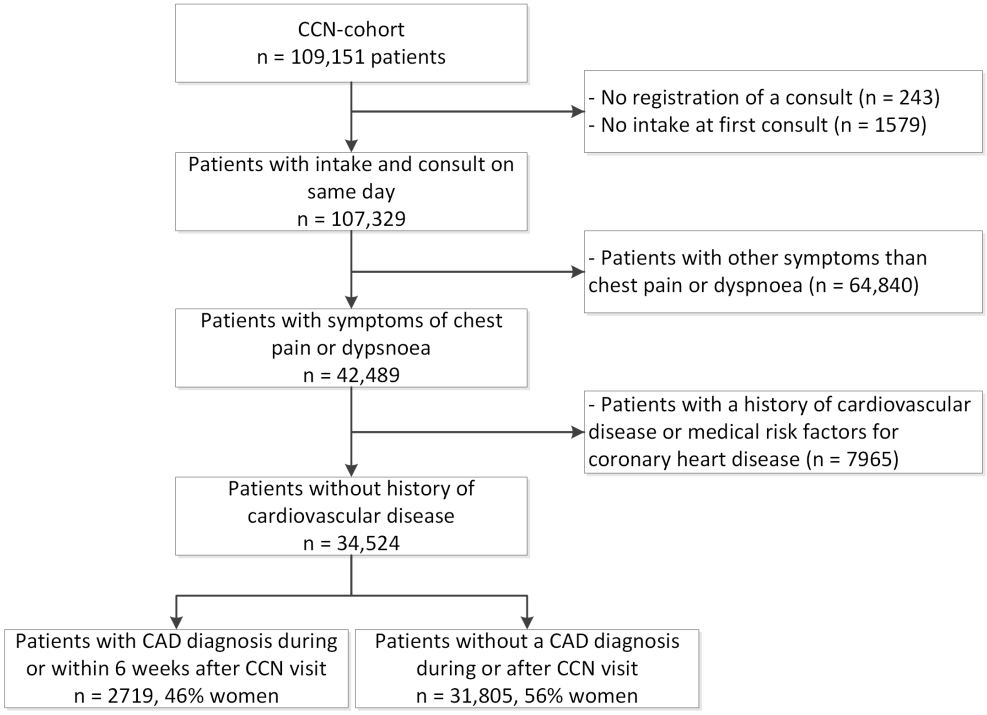


Figure 1 Flowchart of patient selection. CCN: Cardiology Centers of the Netherlands, CAD: coronary artery disease

Missing values

Features with only missing values in either man or women were removed (n = 13). Variables with >80% missing values in either males or females were replaced with a dummy variable that indicated missingness. The original feature was removed from the dataset (n = 94). A missing dummy was created for all lab features, as missingness in measured lab values can be an important source of information.[25] After this, the remaining features with missing values (n = 183) were filled with random sample imputation to retain the original distribution of the feature.[26]

Model development

A logistic regression (LR) with lasso feature selection and a gradient boosting model were trained to identify whether PTP could be improved with the inclusion of diagnostic features. For both types of ML algorithms, three different models were trained; one on a dataset that included men and women, one on a dataset with only men and one with only women. Each dataset was divided into a set for training (75%) and testing (25%). Stratification on a diagnosis of CAD was done to ensure a similar percentage of patients with a diagnosis in the training and test datasets.

LOGISTIC REGRESSION WITH LASSO FEATURE SELECTION

Feature selection and estimation of the coefficients was done with lasso LR. Lasso LR models were developed with the glmnet package in RStudio (RStudio: Integrated Development Environment for R, http://www.rstudio.com/, version 1.3.1093) with the R language (R Foundation for Statistical Computing, https://ww.R-project.org , version 4.0.2).

GRADIENT BOOSTING ALGORITHM

The xgboost package was used to develop a gradient boosting model for identification of CAD. The pipeline for development of the optimal model consisted of different steps, which is elaborated upon in the methods section of the supplemental materials. The gradient boosting algorithm was developed and trained in Python (Python Software Foundation, https://www.python.org, version 3.7.9).

Statistical analysis

Descriptive statistics of the included population are presented as mean with standard deviation (SD) or median with interquartile range (IQR), where appropriate, for continuous variables and counts and percentages for categorical variables. The distribution of men and women with and without a diagnosis of CAD was evaluated in three different categories (<0.05; no diagnostic testing required, 0.05-0.15; non-invasive diagnostic testing may be considered, >0.15; non-invasive diagnostic testing is most beneficial[4]) and visualized with a box plot.

The models; PTP, Lasso LR and gradient boosting algorithm, were compared using differ-

ent metrics. Binary outcome of the models was based on the calculated probability. To determine the classification (CAD or no CAD), two different cut-off values were used; 0.05 and 0.15. These cut-off values were based upon the current guidelines.[4,5] The metrics and their calculations are shown in Supplementary table 2. Confidence intervals around the metrics were obtained with bootstrapping (n=500) on the complete dataset. New Lasso LR and gradient boosting models were trained in each bootstrap.
The TRIPOD statement for the reporting of prognostic and diagnostic models was, where appropriate, followed.[27]

Net reclassification Index for comparison of pre-test probability and machine learning

For the analysis of the categorization of patients suspected of CAD, the different categories as defined for the PTP were used; <0.05, 0.05-0.15 and >0.15. A comparison was made between patient categorization based on the PTP and the Lasso LR and between PTP and gradient boosting.[4,5] This was done with category-based net reclassification index (NRI)[28], with the PTP as the reference value for classification of CAD. The calculated metrics were event NRI, non-event NRI and overall NRI. Event NRI is the proportion of patients with CAD that are correctly classified into a higher risk category based on the new model, whereas the non-event NRI is the proportion of patients without CAD that are correctly classified into a lower risk category with the new model. The overall NRI is the sum of these measures.[28]

RESULTS

Study population

Table 1 summarizes the clinical characteristics of the included patients (n = 34,524, 55.4% women). Overall mean age was 53 (SD: 13.7) years in men and 57 (SD: 13.5) years in women. Individuals with a diagnosis of CAD were overall older; 62 vs. 52 years and 63 vs. 56 years for men and women, respectively and overall had more traditional cardiovascular risk factors (diabetes in men; 12.1% vs. 7.3% and in women; 10.2% vs. 6.8%, hypertension in men; 32.8% vs. 24.4%, and in women; 38.7% vs. 29%, and dyslipidaemia in men; 19.2% vs. 14.2%, and in women; 22.4% vs. 13.3%). Supplementary table 4 shows the distribution of baseline variables between training and test set. No significant differences at baseline were observed between these datasets.

Evaluation of pre-test probability of having obstructive CAD in men and women

Comparison of the pre-test probability in men and women showed that women had a lower median PTP than men (6% [IQR: 3-10] vs. 11% [IQR: 3-22], tested with Wilcoxon rank-sum; $p < 0.01$). This is also shown in Figure 2, where the distribution of the PTP is displayed per sex in individuals with and without a diagnosis of CAD. When investigating

Table 1 Baseline table of included patients, stratified by sex and diagnosis of angina pectoris or coronary

	Men			
	Overall	**No CAD diagnosis**	**CAD Diagnosis**	
n	15,383	13,917	1466	
AGE (mean (SD))	53 (13.7)	52 (13.8)	62 (10.3)	
BMI (mean (SD))	26.8 (4.1)	26.8 (4.1)	26.9 (3.9)	
Classification of BMI (%)				
Normal weight	5168 (34.0)	4710 (34.3)	458 (31.5)	
Obesity class I	2190 (14.4)	1956 (14.2)	234 (16.1)	
Obesity class II	468 (3.1)	442 (3.2)	26 (1.8)	
Obesity class III	153 (1.0)	136 (1.0)	17 (1.2)	
Overweight	7133 (46.9)	6419 (46.7)	714 (49.0)	
Underweight	83 (0.5)	76 (0.6)	7 (0.5)	
SBP (mean (SD))	142.03 (19.16)	141.32 (18.91)	148.80 (20.18)	
DBP (mean (SD))	85.43 (11.66)	85.23 (11.66)	87.40 (11.47)	
Having chest pain complaints (%)	12,997 (84.5)	11,641 (83.6)	1356 (92.5)	
Having complaints of dyspnoea (%)	2599 (16.9)	2463 (17.7)	136 (9.3)	
Smoking status (%)				
Current	6038 (41.8)	5707 (43.6)	331 (24.4)	
Ever	4935 (34.2)	4368 (33.4)	567 (41.9)	
Never	3474 (24.0)	3018 (23.1)	456 (33.7)	
Diabetes (%)	1181 (7.7)	1004 (7.3)	177 (12.1)	
Hypertension (%)	3856 (25.2)	3375 (24.4)	481 (32.8)	
Dyslipidaemia (%)	2253 (14.7)	1971 (14.2)	282 (19.2)	
Family history of atherosclerosis (%)				
Negative	4685 (44.3)	4209 (44.8)	476 (39.6)	
Unknown	1285 (12.1)	1112 (11.8)	173 (14.4)	
Positive	4617 (43.6)	4065 (43.3)	552 (46.0)	
Chest pain (%)				
Non-anginal	3415 (60.7)	3301 (68.0)	114 (14.7)	
Atypical	771 (13.7)	658 (13.6)	113 (14.6)	
Typical	1444 (25.6)	895 (18.4)	549 (70.7)	
10-year SCORE CVD (median (IQR))	2.20 [0.70, 5.53]	2.02 [0.63, 5.14]	4.80 [2.32, 9.24]	
PTP (median (IQR))	11 [3, 22]	11 [3, 22]	24 [17, 32]	
PTP category (%)				
<0.05	4994 (32.5)	4897 (35.2)	97 (6.6)	
0.05-0.15	3594 (23.4)	3352 (24.1)	242 (16.5)	
>0.15	6795 (44.2)	5668 (40.7)	1127 (76.9)	

BMI: Body mass index, SBP: systolic blood pressure, DBP: diastolic blood pressure, SCORE CVD: Systematic

artery disease.

		Women			
	Missing	Overall	No CAD diagnosis	CAD Diagnosis	Missing
		19,141	17,888	1253	
	0.0	57 (13.5)	56 (13.5)	63 (10.6)	0.0
	1.2	26.6 (5.2)	26.6 (5.3)	26.9 (4.9)	1.1
	1.2				1.1
		7979 (42.2)	7502 (42.4)	477 (38.3)	
		2911 (15.4)	2698 (15.3)	213 (17.1)	
		990 (5.2)	923 (5.2)	67 (5.4)	
		410 (2.2)	389 (2.2)	21 (1.7)	
		6362 (33.6)	5914 (33.5)	448 (36.0)	
		272 (1.4)	254 (1.4)	18 (1.4)	
	1.3	139.45 (21.93)	138.89 (21.81)	147.50 (22.13)	1.4
	1.2	83.73 (11.83)	83.59 (11.82)	85.68 (11.79)	1.3
	0.0	15,469 (80.8)	14,340 (80.2)	1129 (90.1)	0.0
	0.0	4062 (21.2)	3912 (21.9)	150 (12.0)	0.0
	6.1				7.2
		6994 (39.4)	6731 (40.5)	263 (23.0)	
		5442 (30.6)	5057 (30.4)	385 (33.7)	
		5329 (30.0)	4835 (29.1)	494 (43.3)	
	0.6	1341 (7.0)	1213 (6.8)	128 (10.2)	0.5
	0.5	5658 (29.7)	5173 (29.0)	485 (38.7)	0.4
	0.5	2658 (13.9)	2377 (13.3)	281 (22.4)	0.4
	31.2				28.7
		4838 (35.5)	4521 (35.8)	317 (31.5)	
		1665 (12.2)	1508 (11.9)	157 (15.6)	
		7140 (52.3)	6608 (52.3)	532 (52.9)	
	63.4				66.0
		4077 (62.6)	3974 (67.0)	103 (17.6)	
		933 (14.3)	857 (14.5)	76 (13.0)	
		1505 (23.1)	1099 (18.5)	406 (69.4)	
	30.1	1.20 [0.28, 3.95]	1.12 [0.26, 3.77]	2.71 [0.94, 6.85]	33.0
	0.0	6 [3, 10]	6 [3, 10]	10 [6, 14]	0.0
	0.0				0.0
		8889 (46.4)	8615 (48.2)	274 (21.9)	
		9340 (48.8)	8644 (48.3)	696 (55.5)	
		912 (4.8)	629 (3.5)	283 (22.6)	

COronary Risk Evaluation[34], PTP: pre-test probability, SD: standard deviation, IQR: interquartile range.

the distribution of CAD per category of the PTP, 21.9% (n=274) of women with a diagnosis of CAD were classified in the lowest category of the PTP, versus only 6.6% (n=92) of men with a diagnosis of CAD (Table 1).

Performance of PTP, Lasso LR and Gradient Boosting models

Table 2 displays the model performance of the different models based on the PTP, general Lasso LR and gradient boosting. The AUC of the PTP in the general models were on average low, ranging from 0.60 to 0.69, suggesting poor performance of the PTP. The AUCs were in general higher in the Lasso LR and gradient boosting models, ranging from 0.71 to 0.80. In the general models tested in men and women combined (Table 2A), the cut-off value of 0.05 resulted in a higher AUC than the 0.15 cut-off for Lasso LR and gradient boosting. When the general models were tested on men only, the AUCs were not different using a 0.05 and 0.15 cut-off (Lasso LR: 0.79 vs 0.80 and gradient boosting: 0.79 vs 0.78, for a 0.05 and 0.15 cut-off, respectively). This also applied to the male-specific models (Table 2B, Lasso LR: 0.77 vs 0.78 and gradient boosting: 0.79 vs 0.80, for a 0.05 and 0.15 cut-off, respectively). The variable cut-off value resulted in different AUC for the PTP in men, albeit this AUC was significantly lower than the AUC of the Lasso LR and gradient boosting (AUC PTP general model, general model: 0.65 vs 0.68, male-specific model: 0.64 vs 0.67, for a 0.05 and 0.15 cut-off, respectively). For women, the 0.05 cut-off value resulted in higher AUC in the general model (general model, AUC PTP: 0.64 vs 0.60, Lasso

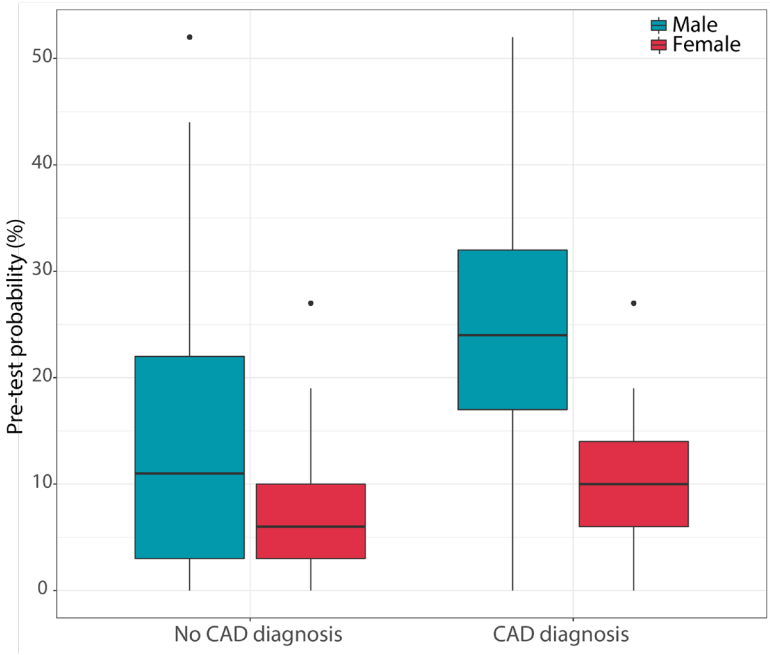


Figure 2 Distribution of pre-test probability in men and women, stratified for a diagnosis of CAD.

Table 2 Performance metrics of the general model, trained on men and women and tested on the complete test population and on men and women from the test population separately, performance metrics of the male-specific and female-specific model. 0.05 and 0.15 indicate the specific cut-off values at which an individual is classified as having CAD.

		Recall/Sensitivity	Specificity	AUC
General model	PTP 0.05	0.87 (0.84,0.89)	0.43 (0.42,0.44)	0.65 (0.63,0.66)
	PTP 0.15	0.50 (0.49,0.55)	0.81 (0.79,0.81)	0.65 (0.64,0.68)
	Lasso LR 0.05	0.88 (0.87,0.92)	0.67 (0.65,0.68)	0.78 (0.77,0.79)
	Lasso LR 0.15	0.61 (0.60,0.66)	0.91 (0.90,0.91)	0.76 (0.75,0.79)
	Gradient boosting 0.05	0.84 (0.83,0.88)	0.71 (0.69,0.72)	0.78 (0.77,0.79)
	Gradient boosting 0.15	0.59 (0.58,0.65)	0.91 (0.90,0.92)	0.75 (0.75,0.78)
	Performance in men			
	PTP 0.05	0.94 (0.91,0.95)	0.35 (0.34,0.37)	0.65 (0.63,0.65)
	PTP 0.15	0.77 (0.73,0.81)	0.60 (0.58,0.61)	0.68 (0.66,0.70)
	Lasso LR 0.05	0.93 (0.90,0.95)	0.65 (0.63,0.67)	0.79 (0.77,0.80)
	Lasso LR 0.15	0.71 (0.66,0.75)	0.89 (0.88,0.90)	0.80 (0.78,0.82)
	Gradient boosting 0.05	0.88 (0.86,0.93)	0.69 (0.67,0.71)	0.79 (0.78,0.81)
	Gradient boosting 0.15	0.68 (0.66,0.76)	0.89 (0.88,0.90)	0.78 (0.78,0.83)
	Performance in women			
	PTP 0.05	0.79 (0.74,0.82)	0.48 (0.47,0.50)	0.64 (0.61,0.65)
	PTP 0.15	0.23 (0.18,0.26)	0.97 (0.96,0.97)	0.60 (0.57,0.61)
	Lasso LR 0.05	0.84 (0.81,0.89)	0.69 (0.65,0.70)	0.76 (0.75,0.78)
	Lasso LR 0.15	0.51 (0.50,0.60)	0.92 (0.91,0.93)	0.71 (0.71,0.76)
	Gradient boosting 0.05	0.79 (0.77,0.85)	0.73 (0.70,0.74)	0.76 (0.74,0.78)
	Gradient boosting 0.15	0.50 (0.46,0.57)	0.94 (0.92,0.93)	0.72 (0.69,0.75)
Male-specific model	PTP 0.05	0.93 (0.91,0.95)	0.35 (0.34,0.36)	0.64 (0.63,0.66)
	PTP 0.15	0.75 (0.73,0.80)	0.59 (0.58,0.61)	0.67 (0.66,0.70)
	Lasso LR 0.05	0.92 (0.90,0.95)	0.62 (0.60,0.65)	0.77 (0.76,0.79)
	Lasso LR 0.15	0.68 (0.66,0.75)	0.89 (0.87,0.90)	0.78 (0.77,0.82)
	Gradient boosting 0.05	0.89 (0.87,0.93)	0.69 (0.65,0.71)	0.79 (0.77,0.80)
	Gradient boosting 0.15	0.71 (0.66,0.75)	0.89 (0.88,0.90)	0.80 (0.78,0.82)
Female-specific model	PTP 0.05	0.81 (0.73,0.81)	0.49 (0.47,0.50)	0.65 (0.61,0.65)
	PTP 0.15	0.24 (0.19,0.27)	0.97 (0.96,0.97)	0.60 (0.58,0.62)
	Lasso LR 0.05	0.86 (0.82,0.89)	0.67 (0.63,0.68)	0.76 (0.74,0.77)
	Lasso LR 0.15	0.49 (0.44,0.53)	0.93 (0.92,0.94)	0.71 (0.69,0.73)
	Gradient boosting 0.05	0.80 (0.78,0.87)	0.75 (0.69,0.75)	0.77 (0.76,0.79)
	Gradient boosting 0.15	0.55 (0.47,0.58)	0.93 (0.92,0.94)	0.74 (0.70,0.75)

AUC: Area Under the receiver-operating Curve, LR: Logistic Regression, PTP: Pre-test probability.

LR: 0.76 vs 0.71, gradient boosting: 0.76 vs 0.72, for a 0.05 and a 0.15 cut-off, respectively) and in the female-specific model (AUC PTP: 0.65 vs 0.60, Lasso LR: 0.76 vs 0.71, gradient boosting: 0.77 vs 0.74, for a 0.05 and a 0.15 cut-off, respectively).
Furthermore, these results showed that models don't have to be developed per sex to improve the classification of patients with CAD (male- and female-specific models in Table 2). The AUC, as described above, sensitivity (in men, 0.05 cut-off value Lasso LR: 0.93 vs 0.92, gradient boosting: 0.88 vs 0.89 and 0.15 cut-off value Lasso LR: 0.71 vs 0.68, gradient boosting: 0.68 vs 0.71 for, respectively, the general model in men and the male-specific model, in women, 0.05 cut-off value Lasso LR: 0.84 vs 0.86, gradient boosting: 0.79 vs 0.80 and 0.15 cut-off value Lasso LR: 0.51 vs 0.49, gradient boosting: 0.50 vs 0.55 for, respectively, the general model in women and the female-specific model) and specificity (in men, 0.05 cut-off value Lasso LR: 0.65 vs 0.62, gradient boosting: 0.69 vs 0.69 and 0.15 cut-off value Lasso LR: 0.89 vs 0.89, gradient boosting: 0.89 vs 0.89 for, respectively, the general model in men and the male-specific model, in women, 0.05 cut-off values Lasso LR: 0.69 vs 0.67, gradient boosting: 0.73 vs 0.75 and 0.15 cut-off value, Lasso LR: 0.92 vs 0.93, gradient boosting: 0.94 vs 0.93 for, respectively, the general model in women and the female-specific model) of the sex-specific models were all within the confidence intervals of the performance of the general model evaluated for each sex specifically.
Sensitivity is an important measure to evaluate as women with CAD are often classified in a low category according to the PTP. This is also illustrated by the sensitivity of the general PTP model with a cut-off of 0.05. In women, sensitivity was 0.79 (95% CI: 0.74-0.82) versus 0.94 (95% CI: 0.91-0.95) in men. Sensitivity significantly improved for women with the general LR model with lasso feature selection model and a 0.05 cut-off value (0.84, 95% CI: 0.81-0.89), while also increasing specificity (0.69, 95% CI: 0.65-0.70) compared to the PTP (0.48, 95% CI: 0.47-0.50). The gradient boosting algorithm did not outperform the Lasso LR in terms of sensitivity in women (0.79, 95% CI: 0.77-0.85) and specificity (0.73, 95% CI: 0.70-0.74). Supplementary table 3 (A, B, C) gives an overview of the models on all different metrics.

Categorization and Reclassification with Machine Learning

Table 3 shows the distribution of patients of the test set in the different categories of the Lasso LR and gradient boosting model compared to the classification with PTP. The net reclassification per model is shown in Table 4A and Table 4B. These tables show that with both the gradient boosting algorithm and the Lasso LR, a high proportion (27%, 95% CI: 27%-40% for Lasso LR and 21%, 95% CI: 20%-34% for gradient boosting) of women with CAD were reclassified into a higher category (event NRI). Event NRI was highest (27%) for women with Lasso LR. Application of the Lasso LR in women also led to an improved classification of non-events by 16% (95% CI: 13%-18%), indicating fewer false positive

findings. In men, the application of the Lasso LR resulted in a small decrease (-6%, 95% CI: -12%-0%) in the reclassification of individuals with CAD. However, the improvement of classification of non-events was 41% (95% CI: 40%-44%), indicating that 41% of men without CAD were reclassified into a lower category and are thus less likely to be exposed to unnecessary diagnostic testing. Sex-specific models (NRI for general Lasso LR tested in men: 0.35, 95% CI: 0.31-0.42, and male-specific Lasso LR: 0.34, 95% CI: 0.29-0.39, NRI for general Lasso LR tested in women: 0.44, 95% CI: 0.42-0.55, and female-specific Lasso LR: 0.39, 95% CI: 0.37-0.50) were not required to improve the NRI in both sexes.

Feature importance

In the bootstrapping analysis, the Lasso LR selected on average 101 (SD: 15), 72 (SD: 13) and 49 (SD: 10) features in the general, male-specific and female-specific model, respectively. Thirty-three features were selected in every general model versus 24 in the male-specific model and 18 in the female-specific model. Overlapping features (14) between these were age, missing text about overall status of the patient, typical chest pain and chest pain characteristics i.e. pressure, radiation, duration, provocative and alleviating factors, abnormal conclusion of stress ECG or myocardial infarction during stress ECG, coronary dysfunction during rest ECG, missing value of troponin, systolic blood pressure and patient number. A complete overview of features and how often these were included can be found in Supplementary table 5.

For the gradient boosting model, only the 30 highest scoring features were selected per model, as analyses (Supplementary figure 2) showed that selection of more than 30 features in the model did not result in an increased improvement of the AUC or other metrics. There were 19 overlapping features selected in the general, male-specific and female-specific model, including typical chest pain, age, chest pain or dyspnoea complaints, chest pain characterizations, i.e. provocation, alleviation, pressure and duration, whether stress ECG was performed, current smoker, units alcohol per day, abnormal conclusion of stress ECG or myocardial infarction, dyspnoea as reason to stop the ECG, missing text about overall status of the patient, missing value for palpation of spleen, for auscultation, for carotid artery upstroke and for ictus cordis.

DISCUSSION

This study showed that in symptomatic women with CAD referred to a CCN center for the first time, the PTP according to the latest ESC guidelines[4,5] for CAD was low, and therefore resulted in a large proportion of false negative results. Approximately 20% of women with a diagnosis of CAD, as defined by the cardiologists, score below the threshold for diagnostic imaging. We showed that inclusion of a large array of diagnostic features from a regular care database in a Lasso LR improved risk stratification, especially in women. Nonetheless, sophisticated AI models, i.e. gradient boosting algorithms, were not re-

Table 3 Distribution of the risk classification in the test set. These tables show the comparison of the classification between the PTP and Lasso LR and the PTP and gradient boosting in patients with and without a diagnosis of CAD. Values are displayed as number of patients, followed by percentage (%) of total with diagnosis or without diagnosis. Red cells are incorrect reclassifications

Patients with a diagnosis of CAD				
		PTP < 0.05	**PTP 0.05-0.15**	**PTP > 0.15**
	Lasso LR < 0.05	35 (5.1)	34 (5)	10 (1.5)
	Lasso LR 0.05-0.15	30 (4.4)	102 (15)	52 (7.6)
	Lasso LR > 0.15	27 (4)	109 (16)	281 (41.3)
	Distribution in men of test set			
	Lasso LR < 0.05	3 (0.9)	12 (3.5)	10 (2.9)
	Lasso LR 0.05-0.15	9 (2.6)	23 (6.6)	42 (12.1)
	Lasso LR > 0.15	10 (2.9)	24 (6.9)	214 (61.7)
	Distribution in women of test set			
	Lasso LR < 0.05	32 (9.6)	22 (6.6)	0 (0)
	Lasso LR 0.05-0.15	21 (6.3)	79 (23.7)	10 (3)
	Lasso LR > 0.15	17 (5.1)	85 (25.5)	67 (20.1)

Patients without a diagnosis of CAD				
		PTP < 0.05	**PTP 0.05-0.15**	**PTP > 0.15**
	Lasso LR < 0.05	2808 (35.3)	1870 (23.5)	636 (8)
	Lasso LR 0.05-0.15	531 (6.7)	842 (10.6)	521 (6.6)
	Lasso LR > 0.15	80 (1)	273 (3.4)	390 (4.9)
	Distribution in men of test set			
	Lasso LR < 0.05	1048 (30.1)	582 (16.7)	620 (17.8)
	Lasso LR 0.05-0.15	162 (4.7)	218 (6.3)	460 (13.2)
	Lasso LR > 0.15	25 (0.7)	57 (1.6)	311 (8.9)
	Distribution in women of test set			
	Lasso LR < 0.05	1760 (39.4)	1288 (28.8)	16 (0.4)
	Lasso LR 0.05-0.15	369 (8.3)	624 (14)	61 (1.4)
	Lasso LR > 0.15	55 (1.2)	216 (4.8)	79 (1.8)

PTP: pre-test probability, LR: Logistic regression

quired for this improvement as a Lasso LR model performed equally well. Furthermore, sex-specific models for classification of CAD did not perform better than a model trained on data that included both sexes.

Strengths of this study are the use of a large study population that represents the actual population for which the PTP has been designed and is intended: patients in regular care with cardiovascular complaints that visit a physician. This is in contrast to the pooled population of two registries[29,30] and one clinical trial[31] in which the PTP was developed. With the use of regular care data, we eliminate the healthy volunteer effect present in study populations of clinical trials[32,33]. The prevalence of CAD in our study population

and indicate a down-classification in case of patients with a diagnosis of CAD and an up-classification in case of patients without a diagnosis of CAD. Green cells are correct reclassifications and indicate an up-classification in case of patients with a diagnosis of CAD and a down-classification in case of patients without a diagnosis of CAD.

		PTP < 0.05	PTP 0.05-0.15	PTP > 0.15
Patients with a diagnosis of CAD	Gradient boosting < 0.05	40 (5.9)	52 (7.6)	17 (2.5)
	Gradient boosting 0.05-0.15	24 (3.5)	88 (12.9)	58 (8.5)
	Gradient boosting > 0.15	28 (4.1)	105 (15.4)	268 (39.4)
	Distribution in men of test set			
	Gradient boosting < 0.05	8 (2.3)	15 (4.3)	17 (4.9)
	Gradient boosting 0.05-0.15	5 (1.4)	19 (5.5)	48 (13.8)
	Gradient boosting > 0.15	9 (2.6)	25 (7.2)	201 (57.9)
	Distribution in women of test set			
	Gradient boosting < 0.05	32 (9.6)	37 (11.1)	0 (0)
	Gradient boosting 0.05-0.15	19 (5.7)	69 (20.7)	10 (3)
	Gradient boosting > 0.15	19 (5.7)	80 (24)	67 (20.1)
		PTP < 0.05	**PTP 0.05-0.15**	**PTP > 0.15**
Patients without a diagnosis of CAD	Gradient boosting < 0.05	2993 (37.6)	1963 (24.7)	699 (8.8)
	Gradient boosting 0.05-0.15	353 (4.4)	774 (9.7)	481 (6)
	Gradient boosting > 0.15	73 (0.9)	248 (3.1)	367 (4.6)
	Distribution in men of test set			
	Gradient boosting < 0.05	1145 (32.9)	600 (17.2)	666 (19.1)
	Gradient boosting 0.05-0.15	69 (2)	189 (5.4)	416 (11.9)
	Gradient boosting > 0.15	21 (0.6)	68 (2)	309 (8.9)
	Distribution in women of test set			
	Gradient boosting < 0.05	1848 (41.4)	1363 (30.5)	33 (0.7)
	Gradient boosting 0.05-0.15	284 (6.4)	585 (13.1)	65 (1.5)
	Gradient boosting > 0.15	52 (1.2)	180 (4)	58 (1.3)

(7.9%) was lower than in the pooled analysis (14.9%), which might be explained by the healthy population in our study, as indicated by the 10-year CVD SCORE[34] (Table 1) and when comparing comorbidities, hypertension: 25% vs 55%, diabetes: 8% vs 16%, dyslipidaemia: 15% vs 60%, in respectively, our population and the pooled population on which the PTP was developed.[5] As the CCN is positioned between the general practitioner (GP) and the hospital cardiologist, the GP has a low threshold to refer patients, which might result in a relatively healthy population. The large study population also enabled sex-specific and sex-stratified analyses. Furthermore, the use of regular care data created the opportunity to use a large array of easily obtainable diagnostic features to identify

Table 4 The event NRI, non-event NRI and overall NRI using the different models on top of the pre-test probability. The upper table shows the NRI of the general models trained on men and women and the lower table displays the NRI of the male- and female-specific models. Three categories were taken into account and reclassification was observed as a change in category using the Lasso logistic regression or the gradient boosting model compared to the pre-test probability.

		Event NRI	Non-event NRI	Overall NRI
General model	Lasso LR	0.10 (0.08, 0.17)	0.27 (0.25,0.29)	0.37 (0.34,0.44)
	Gradient boosting	0.04 (0.04, 0.13)	0.31 (0.29,0.33)	0.35 (0.35,0.44)
	NRI in men of test set			
	Lasso LR	-0.06 (-0.12, 0.00)	0.41 (0.40,0.44)	0.35 (0.31,0.42)
	Gradient boosting	-0.12 (-0.12,-0.02)	0.44 (0.43,0.47)	0.32 (0.33,0.43)
	NRI in women of test set			
	Lasso LR	0.27 (0.27, 0.40)	0.16 (0.13,0.18)	0.44 (0.42,0.55)
	Gradient boosting	0.21 (0.20, 0.34)	0.21 (0.18,0.22)	0.42 (0.40,0.53)
		Event NRI	**Non-event NRI**	**Overall NRI**
Sex-specific models	**Male-specific model**			
	Lasso LR	-0.07 (-0.11,-0.01)	0.41 (0.38,0.43)	0.34 (0.29,0.39)
	Gradient boosting	-0.05 (-0.13,-0.02)	0.45 (0.42,0.46)	0.40 (0.31,0.42)
	Female-specific model			
	Lasso LR	0.25 (0.23, 0.36)	0.14 (0.12,0.17)	0.39 (0.37,0.50)
	Gradient boosting	0.23 (0.21, 0.34)	0.22 (0.19,0.24)	0.46 (0.43,0.56)

LR: Logistic regression, NRI: Net Reclassification Index

features that might be important for diagnostic risk stratification of CAD. This large array of available features and the use of a large population made this dataset also very applicable for the use of sophisticated AI to identify the added value of such models in clinical practice. On top, the use of a large feature set enabled thorough investigation of the added value of cardiovascular diagnostic screening centers in the work-up of patients with cardiovascular complaints. Additional diagnostic screening in these centers can improve the risk-stratification of patients with cardiovascular complaints and decrease the number of false positive patients that are referred for diagnostic imaging, thereby reducing costs and radiation burden to the patient. Nonetheless, diagnostic screening inevitably results in a number of false positives, as the focus of screening is on elimination of false negative results.[35]

Limitations of this study include the use of a diagnosis of AP or CAD as the primary outcome to be modelled. It is preferred to use the outcome of a diagnostic test, used as a ground truth, i.e. invasive CAG or CCTA in case of a diagnosis of CAD. Unfortunately, diagnostic screening information, as invasive CAG or CCTA, was not available for most of the population. The use of a registered diagnosis of CAD as the outcome in our main

analysis might have led to an overestimation of events. During data collection, previous guidelines were still in practice.[36] These guidelines overestimated the probability of CAD.[6,7] Even so, cardiologists could have set an initial working diagnosis, based on these guidelines, to initiate treatment or refer for diagnostics. This might have led to an increased number of false positives.

This study was furthermore limited by the data that was used to validate the model, as this data was obtained from the same database used for training. To perform a true validation of the models developed, an external validation dataset would be more appropriate. However, it is hard to find a comparable dataset that includes all the diagnostic features in this regular care database, due to the one-stop shop design of the CCN. This one-stop shop is unique as it includes a complete cardiac diagnostic work-up, including basic lab variables, stress and rest ECG and an echocardiography. This diagnostic pathway is not available at the GP and is too elaborate to be performed in a general hospital setting. To develop a more generalizable model, we could have chosen to only include baseline features that could be obtained from GP records, instead of all diagnostic features. However, the current set-up of this study provided us with more information about important diagnostic variables for CAD, besides general baseline characteristics, and with possible differences between men and women in presentation of CAD. Furthermore, it showed that the one-stop shop diagnostic work-up is a good set-up for screening of patients with cardiovascular symptoms, as features from the stress ECG were often selected in the Lasso LR and the gradient boosting model.

In this study we showed the performance of the PTP for CAD and improvement of this diagnostic tool with Lasso LR and ensemble boosting. We have shown that women are underdiagnosed in this dataset with the current cut-off values for the PTP. A similar trend was seen by Winther et al.[7] In this study the percentage of women with CAD in the lowest category of the PTP was approximately 10%. Also in the study by Bing et al.[6] the PTP slightly underestimated the prevalence of CAD. It could be hypothesized that women and men might benefit from a different cut-off value of the PTP, which we have evaluated in this study as the 0.05 and the 0.15 cut-off value. Our results showed that indeed a cut-off of 0.05 in women led to significantly higher AUCs and sensitivity than a cut-off of 0.15, whereas for men AUCs of the PTP were more or less similar for 0.05 and 0.15 cut-off. On top, our results confirm the patient-specific evaluation of additional risk factors, not incorporated in the PTP, for a better risk stratification. The selected risk factors by the ESC guidelines are outcome of exercise ECG, cardiovascular comorbidities i.e. dyslipidaemia, diabetes, hypertension, smoking and a family history of CVD, resting ECG changes and left ventricular dysfunction.[4] In the comparison of these risk factors to the feature importance in our results, we see some overlap for the inclusion of the results of the ECG stress test, smoking and resting ECG changes. However, in the feature importance of the Lasso

LR, there was also a prominent role for chest pain characterization beyond the characteristics used to classify non-anginal, atypical and typical chest pain.
AI is depicted as the hope of healthcare, although validation of AI tools is limited[37] and widespread implementation in contemporary clinical care is still lacking behind all the efforts made in research[38]. With the presented study, our intention was to evaluate the performance of current risk stratification tools and improvement with AI. AI has shown to perform very well in specific tasks with unstructured data, e.g. cardiovascular imaging[39] or ECGs[40-42]. However, the database used in this study is appropriately structured and might thus not benefit from AI tools as much as unstructured data. This was also shown by the similar performance of Lasso LR models compared to gradient boosting models. More studies were unable to show improved risk-stratification or showed only small improvement on (semi-)structured data analysed with AI and compared to traditional risk models or traditional statistical methods. First, ML models performed more poorly in comparison with traditional statistics in the prediction of outcomes in atrial fibrillation[43], and second, ML models showed only limited increase of the AUC for prediction of obstructive CAD on clinical variables (AUC ML models: 0.773, 95% CI: 0.76-0.79 vs. AUC CAD consortium clinical score: 0.734, 95% CI: 0.717-0.751).[44] An important message, that results from our presented study is that emphasis in the clinical evaluation of patients with suspected CAD should be on a clear and structured anamnesis and identification of risk factors. In that case, Lasso LR might be used to improve risk-stratification of patients with CAD. Nonetheless, it is no easy tool to use, as it incorporates on average 101 features and requires data modification. Furthermore, features derived from ECG stress testing might improve risk stratification, as these features were often selected in our models.

CONCLUSION

This study adds to the validation of the PTP and specifically to the performance of the PTP in patients with chest pain or dyspnoea referred to a cardiac diagnostic screening center. It showed that performance of a sex-age-complaint risk classification tool for diagnosis of CAD is acceptable, but also that a Lasso LR outperformed the PTP and a sophisticated gradient boosting algorithm, when many diagnostic features are used. Reclassification into a different category using Lasso LR was specifically useful in women, as the prevalence of CAD was underestimated by the PTP in low-risk women.

REFERENCES

1. On behalf of the Atlas Writing Group. European Society of Cardiology: Cardiovascular Disease Statistics 2019. *Eur Heart J*. 2019;0:1-74. http://www.ehnheart.org/cvd-statistics.html
2. Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol*. 2017;70(1):1-25. doi:10.1016/j.jacc.2017.04.052
3. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis Mak*. 2021;21(1):1-23. doi:10.1186/s12911-021-01488-9
4. Knuuti J, Wijns W, Achenbach S, et al. 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes. *Eur Heart J*. 2020;41(3):407-477. doi:10.1093/eurheartj/ehz425
5. Juarez-Orozco LE, Saraste A, Capodanno D, et al. Impact of a decreasing pre-test probability on the performance of diagnostic tests for coronary artery disease. *Eur Heart J Cardiovasc Imaging*. 2019;20(11):1198-1207. doi:10.1093/ehjci/jez054
6. Bing R, Singh T, Dweck MR, et al. Validation of European Society of Cardiology pre-test probabilities for obstructive coronary artery disease in suspected stable angina. *Eur Hear J - Qual Care Clin Outcomes*. 2020;6(4):293-300. doi:10.1093/ehjqcco/qcaa006
7. Winther S, Schmidt SE, Rasmussen LD, et al. Validation of the European Society of Cardiology pre-test probability model for obstructive coronary artery disease. *Eur Heart J*. Published online 2020:1-11. doi:10.1093/eurheartj/ehaa755
8. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. *Nature*. 2019;575(7781):137-146. doi:10.1038/s41586-019-1657-6
9. Vogel B, Acevedo M, Appelman Y, et al. The Lancet Commissions The Lancet women and cardiovascular disease Commission : reducing the global burden by 2030. Published online 2021. doi:10.1016/S0140-6736(21)00684-X
10. Genders TSS, Steyerberg EW, Hunink MGM, et al. Prediction model to estimate presence of coronary artery disease: Retrospective pooled analysis of existing cohorts. *BMJ*. 2012;344(7862):1-13. doi:10.1136/bmj.e3485
11. Oei HHS, Vliegenthart R, Hak AE, et al. The association between coronary calcification assessed by electron beam computed tomography and measures of extracoronary atherosclerosis: The Rotterdam Coronary Calcification Study. *J Am Coll Cardiol*. 2002;39(11):1745-1751. doi:10.1016/S0735-1097(02)01853-3
12. Blankstein R, Ahmed W, Bamberg F, et al. Comparison of exercise treadmill testing with cardiac computed tomography angiography among patients presenting to the emergency room with chest pain: The rule out myocardial infarction using computer-assisted tomography (ROMICAT) study. *Circ Cardiovasc Imaging*. 2012;5(2):233-242. doi:10.1161/CIRCIMAGING.111.969568
13. Salokari E, Laukkanen JA, Lehtimaki T, et al. The Duke treadmill score with bicycle ergometer: Exercise capacity is the most important predictor of cardiovascular mortality. *Eur J Prev Cardiol*. 2018;22:1-9. doi:10.1177/2047487318804618
14. Ferrari R, Abergel H, Ford I, et al. Gender- and age-related differences in clinical presentation and management of outpatients with stable coronary artery disease. *Int J Cardiol*. 2013;167(6):2938-2943. doi:10.1016/j.ijcard.2012.08.013
15. Groepenhoff F, Eikendal ALM, Charlotte Onland-Moret N, et al. Coronary artery disease prediction in women and men using

chest pain characteristics and risk factors: An observational study in outpatient clinics. *BMJ Open*. 2020;10(4). doi:10.1136/bmjopen-2019-035928
16. George J, Rapsomaniki E, Pujades-Rodriguez M, et al. How does cardiovascular disease first present in women and men? *Circulation*. 2015;132(14):1320-1328. doi:10.1161/CIRCULATIONAHA.114.013797
17. Lee HC, Park JS, Choe JC, et al. Prediction of 1-Year Mortality from Acute Myocardial Infarction Using Machine Learning. *Am J Cardiol*. 2020;133:23-31. doi:10.1016/j.amjcard.2020.07.048
18. Tohyama T, Ide T, Ikeda M, et al. Machine learning-based model for predicting 1 year mortality of hospitalized patients with heart failure. *ESC Hear Fail*. Published online 2021:1-9. doi:10.1002/ehf2.13556
19. Hernesniemi JA, Mahdiani S, Tynkkynen JA, et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome–the MADDEC study. *Ann Med*. 2019;51(2):156-163. doi:10.1080/07853890.2019.1596302
20. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One*. 2018;13(8):1-20. doi:10.1371/journal.pone.0202344
21. Samad MD, Ulloa A, Wehner GJ, et al. Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning. *JACC Cardiovasc Imaging*. 2019;12(4):681-689. doi:10.1016/j.jcmg.2018.04.026
22. Bots SH, Siegersma KR, Onland-Moret NC, et al. Routine clinical care data from thirteen cardiac outpatient clinics: design of the Cardiology Centers of the Netherlands (CCN) database. *BMC Cardiovasc Disord*. 2021;21(1):1-9. doi:10.1186/s12872-021-02020-7
23. Diamond GA. A clinically relevant classification of chest discomfort. *J Am Coll Cardiol*. 1983;1(2):574-575. doi:10.1016/S0735-1097(83)80093-X
24. Punwasi R, Musson REA. What's Lab. UMC Utrecht Laboratoriumbepalingen LKCH/LTI. Accessed October 18, 2021. http://lkch.nl/bepalingen/
25. Abedi V, Li J, Shivakumar MK, et al. Increasing the Density of Laboratory Measures for Machine Learning Applications. *J Clin Med*. 2020;10(1):103. doi:10.3390/jcm10010103
26. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59(10):1087-1091. doi:10.1016/j.jclinepi.2006.01.014
27. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation*. 2015;131(2):211-219. doi:10.1161/CIRCULATIONAHA.114.014508
28. Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net Reclassification Improvement: Computation, Interpretation and Controversies. *Ann Intern Med*. 2014;160:122-131.
29. Reeh J, Therming CB, Heitmann M, et al. Prediction of obstructive coronary artery disease and prognosis in patients with suspected stable angina. *Eur Heart J*. 2019;40(18):1426-1435. doi:10.1093/eurheartj/ehy806
30. Cheng VY, Berman DS, Rozanski A, et al. Performance of the traditional age, sex, and angina typicality-based approach for estimating pretest probability of angiographically significant coronary artery disease in patients undergoing coronary computed tomographic angiography: Results from the mul. *Circulation*. 2011;124(22):2423-2432. doi:10.1161/CIRCULATIONAHA.111.039255
31. Foldyna B, Udelson JE, Karády J, et al.

Pretest probability for patients with suspected obstructive coronary artery disease: Re-evaluating Diamond-Forrester for the contemporary era and clinical implications: Insights from the PROMISE trial. *Eur Heart J Cardiovasc Imaging*. 2019;20(5):574-581. doi:10.1093/ehjci/jey182

32. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am J Epidemiol*. 2007;165(8):874-881. doi:10.1093/aje/kwk075
33. Leening MJG, Heeringa J, Deckers JW, et al. Healthy volunteer effect and cardiovascular risk. *Epidemiology*. 2014;25(3):470-471. doi:10.1097/EDE.0000000000000091
34. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J*. 2003;24(11):987-1003. doi:10.1016/S0195-668X(03)00114-3
35. Dans LF, Silvestre MAA, Dans AL. Trade-off between benefit and harm is crucial in health screening recommendations. Part I: General principles. *J Clin Epidemiol*. 2011;64(3):231-239. doi:10.1016/j.jclinepi.2010.09.009
36. The task force on the management of stable coronary artery disease of the European Society of Cardiology. 2013 ESC guidelines on the management of stable coronary artery disease: *Eur Heart J*. 2013;34:2949-3003. doi:10.1093/eurheartj/eht296
37. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*. 2021;31(6):3797-3804. doi:10.1007/s00330-021-07892-z
38. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *npj Digit Med*. 2018;1(1):10-12. doi:10.1038/s41746-018-0048-y
39. Sermesant M, Delingette H, Cochet H, Jaïs P, Ayache N. Applications of artificial intelligence in cardiovascular imaging. *Nat Rev Cardiol*. 2021;0123456789. doi:10.1038/s41569-021-00527-2
40. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*. 2021;18(7):465-478. doi:10.1038/s41569-020-00503-2
41. Alizadehsani R, Abdar M, Roshanzamir M, et al. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Comput Biol Med*. 2019;111(June):103346. doi:10.1016/j.compbiomed.2019.103346
42. Tan JH, Hagiwara Y, Pang W, et al. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Comput Biol Med*. 2018;94(December 2017):19-26. doi:10.1016/j.compbiomed.2017.12.023
43. Loring Z, Mehrotra S, Piccini JP, et al. Machine learning does not improve upon traditional regression in predicting outcomes in atrial fibrillation: An analysis of the ORBIT-AF and GARFIELD-AF registries. *Europace*. 2020;22(11):1635-1644. doi:10.1093/europace/euaa172
44. Al'Aref SJ, Maliakal G, Singh G, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: Analysis from the CONFIRM registry. *Eur Heart J*. 2020;41(3):359-367. doi:10.1093/eurheartj/ehz565

SUPPLEMENTARY MATERIALS

Supplementary methods

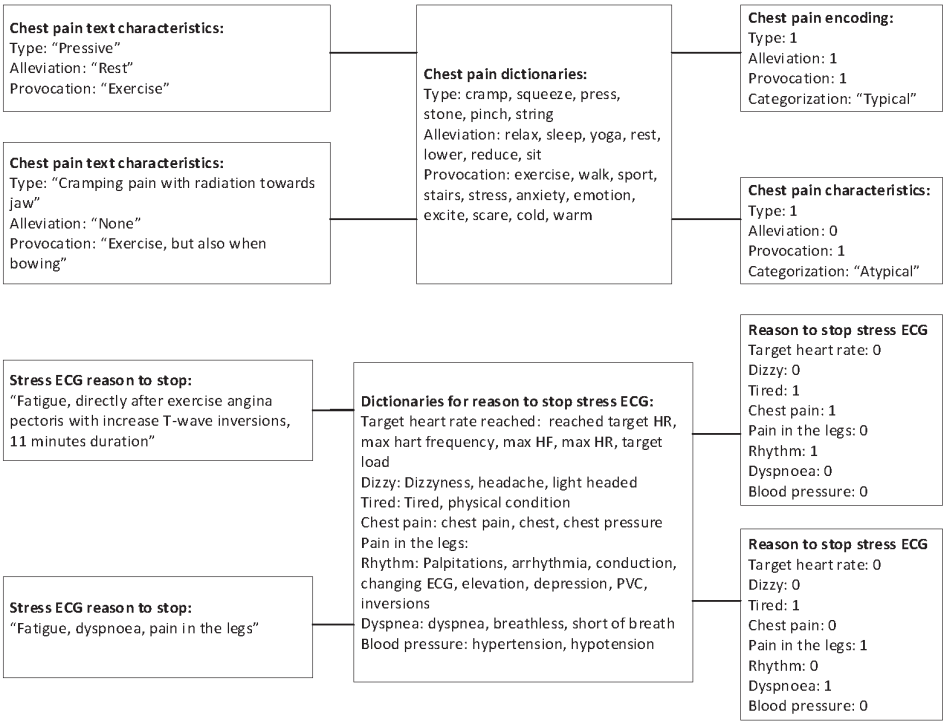
DEVELOPMENT OF GRADIENT BOOSTING MODEL

The development of the gradient boosting algorithm consisted of a series of experiments to develop the optimal gradient boosting model for this dataset. First, an experiment was executed to determine the optimal feature set and number of features to include. Therefore, all features were separated in different classes, as described in Supplementary table 3. The value of each group separately was determined, all groups combined and each group separately combined with the baseline characteristics, resulting in the evaluation of 10 different groups. A 10-fold cross-validation was done on the dataset for training. The optimal feature set was determined through evaluation of the AUC on the different datasets. Second, feature importance was evaluated, followed by a calculation of the optimal number of features to include. Feature importance was obtained from each of the features in the optimal feature set as determined by the first experiment. Another 10-fold cross-validation was done on the train set to determine the average gain per feature across each of the 10 folds. After this, features were ranked according to the average gain and added one-by-one to the model. This resulted in a visualization of the development of the AUC and other performance metrics with increasing features in the model. Based on the optimal value of the AUC, the optimal number of features to include was determined. Third, an elaborate grid search was done to determine the effect of the different model parameters. Model parameters that were optimized in this process were child weight, gamma, subsample, colsample by tree and maximum depth were the parameters optimized in this process. The final version of the model was trained, taking into account all information of the previous experiments, thus; feature groups to include, optimal number of features and optimal settings of the model parameters as derived from the grid search. The final model was trained train dataset and tested on the test set. During the bootstrapping process, these settings were not changed to induce comparability of the models.

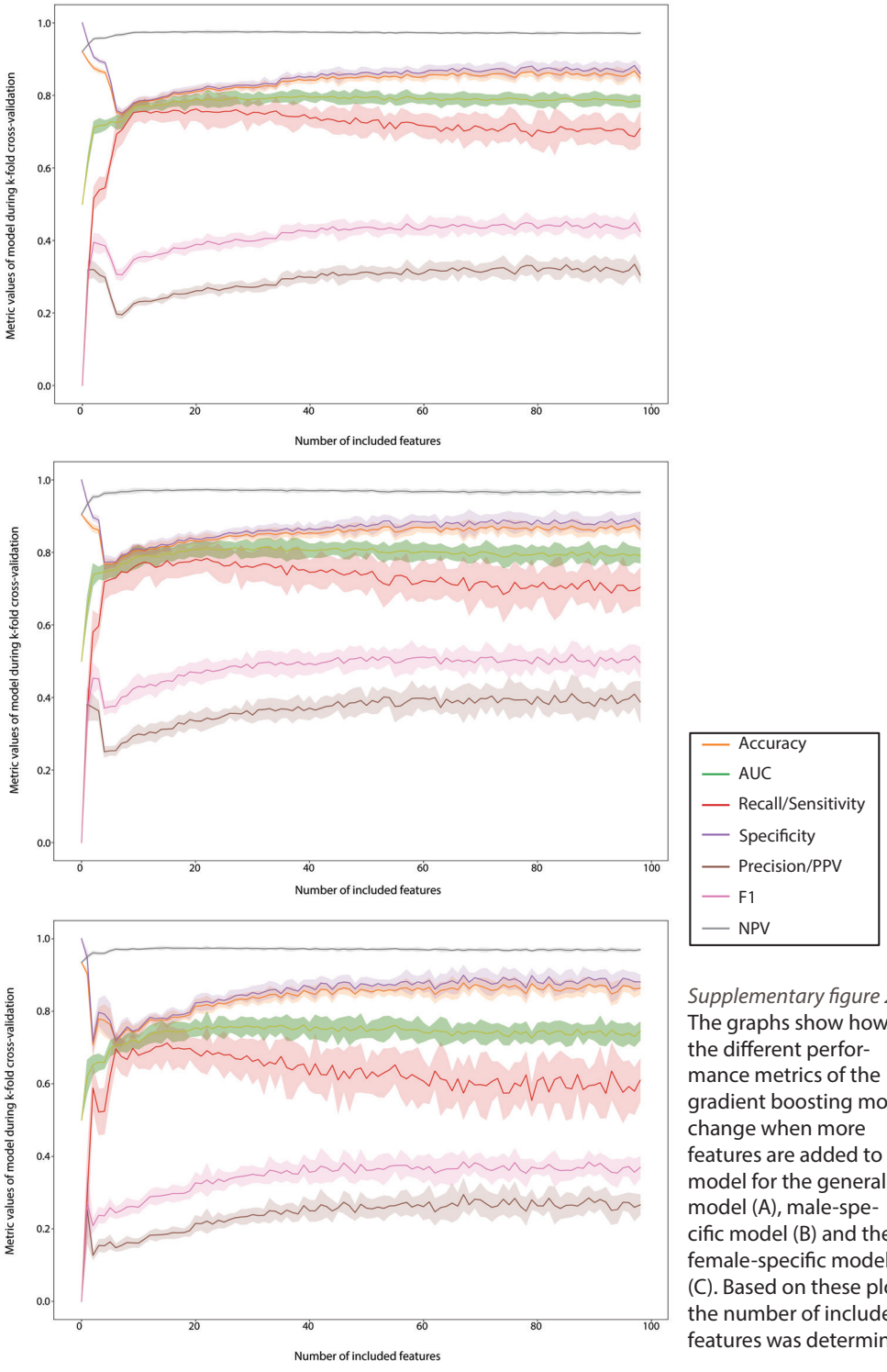
In these experiments, the highest area under the ROC-curve was used as the metric to determine the optimal threshold of the model to classify events and non-events. Early stopping of training of the model on the train dataset was induced to avoid overfitting in the fitting process.

Training and validation of the gradient boosting model was done in Python (Python Software Foundation, https://www.python.org, version 3.7.9), using the packages; scikit-learn and xgboost.

Supplementary figures



Supplementary figure 1 Illustrative example of text retrieval methods for classification of chest pain complaints, reason to stop the stress ECG and other free text features in the EHR.
ECG: electrocardiogram, PVC: premature ventricular contraction, HF: heart frequency, HR: heart rate



Supplementary figure 2
The graphs show how the different performance metrics of the gradient boosting model change when more features are added to the model for the general model (A), male-specific model (B) and the female-specific model (C). Based on these plots, the number of included features was determined.

Supplementary table 1 Groups of features used in the analysis

Category	Variables	Amount of features
Baseline characteristics	Sex, age, comorbidities, risk factors, medical history, family diseases, medications at baseline, chief complaint, physical examination.	176
Exercise ECG	Betablocker use and medication, exercise and stress rhythm and possible arrhythmias, changes in morphology of Q-, ST and T-tops, blood pressure response, indication, stopping reason, conclusion	88
Rest ECG	Presence of ST-changes, pathological Q-changes, negative T-tops, rhythm changes, conduction times and conduction delays, indication, conclusion, atrial dilatation	71
Echocardiogram	Atrial and ventricular dimension, flow measurements, valve observations and insufficiencies, wall motion, ventricular function, ejection fraction, intimal media thickness, diastolic and systolic function.	140
Lab values	HB, potassium, sodium, creatinine, HDL-, LDL- and total cholesterol, triglycerides, glucoses, GFR, LPA, BNP, TSH, ALAT, ASAT, ALP, CRP, CK, D-Dimer, DH, NT-proBNP, GGT, troponin	46

ECG: electrocardiogram, HB: haemoglobin, HDL: high-density lipoprotein, LDL: low-density lipoprotein, GFR: glomerular filtration rate, LPA: lipoprotein (a), BNP: B-type natriuretic peptide, TSH: thyroid-stimulating hormone, ALAT: alanine aminotransferase, ASAT: aspartate aminotransferase, ALP: alkaline phosphatase, CRP: C-reactive protein, CK: creatine kinase, GGT: gamma-glutamyl transferase

Supplementary table 2 Overview of used metrics.

Metrics	Calculation	Explanation
Accuracy	(TP + TN)/(TN + TP + FN + FP)	Proportion of samples that have been correctly classified.
Sensitivity/Recall	TP/(TP + FN)	Proportion of true positives that are classified correctly, given the number of actual positives.
Specificity	TN/(TN + FP)	Proportion of true negatives that are classified correctly, given the number of actual negatives.
Precision/Positive predictive value (PPV)	TP/(TP + FP)	Proportion of true positives that are classified correctly, given the number of positive classifications.
Negative predictive value (NPV)	TN/(TN + FN)	Proportion of true negatives that are classified correctly, given the number of negative classifications.
F1-score	2*((precision*recall)/(precision + recall))	Balance between precision and recall.
Detection rate	TP/(TN + FP + FN + FP)	Proportion of classifications that are correctly positively classified.
Detection prevalence	(TN + FP)/ (TN + TP + FN + FP)	Proportion of classifications that are positively classified
Balanced accuracy	(Sensitivity + Specificity)/2	Average of sensitivity and specificity.

TP: true positives, TN: true negatives, FP: false positives, FN: false negatives

Supplementary table 3 Performance metrics of the general model, trained on men and women and tested on the complete test population and on men and women from the test population sepa-

		Accuracy	Balanced Accuracy	Recall/Sensitivity
General model	PTP 5%	0.46 (0.45,0.47)	0.65 (0.63,0.66)	0.87 (0.84,0.89)
	PTP 15%	0.78 (0.77,0.79)	0.65 (0.64,0.68)	0.50 (0.49,0.55)
	Lasso LR 5%	0.69 (0.67,0.70)	0.78 (0.77,0.79)	0.88 (0.87,0.92)
	Lasso LR 15%	0.88 (0.88,0.89)	0.76 (0.75,0.79)	0.61 (0.60,0.66)
	Gradient boosting 5%	0.72 (0.70,0.73)	0.78 (0.77,0.79)	0.84 (0.83,0.88)
	Gradient boosting 15%	0.89 (0.88,0.89)	0.75 (0.75,0.78)	0.59 (0.58,0.65)
	Performance in men			
	PTP 5%	0.41 (0.40,0.42)	0.65 (0.63,0.65)	0.94 (0.91,0.95)
	PTP 15%	0.62 (0.60,0.62)	0.68 (0.66,0.70)	0.77 (0.73,0.81)
	Lasso LR 5%	0.67 (0.66,0.70)	0.79 (0.77,0.80)	0.93 (0.90,0.95)
	Lasso LR 15%	0.87 (0.87,0.89)	0.80 (0.78,0.82)	0.71 (0.66,0.75)
	Gradient boosting 5%	0.71 (0.70,0.73)	0.79 (0.78,0.81)	0.88 (0.86,0.93)
	Gradient boosting 15%	0.87 (0.87,0.89)	0.78 (0.78,0.83)	0.68 (0.66,0.76)
	Performance in women			
	PTP 5%	0.51 (0.49,0.52)	0.64 (0.61,0.65)	0.79 (0.74,0.82)
	PTP 15%	0.91 (0.91,0.92)	0.60 (0.57,0.61)	0.23 (0.18,0.26)
	Lasso LR 5%	0.70 (0.67,0.71)	0.76 (0.75,0.78)	0.84 (0.81,0.89)
	Lasso LR 15%	0.89 (0.88,0.90)	0.71 (0.71,0.76)	0.51 (0.50,0.60)
	Gradient boosting 5%	0.73 (0.70,0.74)	0.76 (0.74,0.78)	0.79 (0.77,0.85)
	Gradient boosting 15%	0.90 (0.89,0.91)	0.72 (0.69,0.75)	0.50 (0.46,0.57)
		Accuracy	**Balanced Accuracy**	**Recall/Sensitivity**
Male-specific model	PTP 5%	0.40 (0.40,0.42)	0.64 (0.63,0.66)	0.93 (0.91,0.95)
	PTP 15%	0.61 (0.60,0.62)	0.67 (0.66,0.70)	0.75 (0.73,0.80)
	Lasso LR 5%	0.65 (0.63,0.68)	0.77 (0.76,0.79)	0.92 (0.90,0.95)
	Lasso LR 15%	0.87 (0.86,0.88)	0.78 (0.77,0.82)	0.68 (0.66,0.75)
	Gradient boosting 5%	0.71 (0.68,0.72)	0.79 (0.77,0.80)	0.89 (0.87,0.93)
	Gradient boosting 15%	0.87 (0.86,0.88)	0.80 (0.78,0.82)	0.71 (0.66,0.75)
		Accuracy	**Balanced Accuracy**	**Recall/Sensitivity**
Female-specific model	PTP 5%	0.52 (0.49,0.52)	0.65 (0.61,0.65)	0.81 (0.73,0.81)
	PTP 15%	0.92 (0.91,0.92)	0.60 (0.58,0.62)	0.24 (0.19,0.27)
	Lasso LR 5%	0.68 (0.65,0.69)	0.76 (0.74,0.77)	0.86 (0.82,0.89)
	Lasso LR 15%	0.90 (0.89,0.91)	0.71 (0.69,0.73)	0.49 (0.44,0.53)
	Gradient boosting 5%	0.76 (0.70,0.75)	0.77 (0.76,0.79)	0.80 (0.78,0.87)
	Gradient boosting 15%	0.90 (0.89,0.91)	0.74 (0.70,0.75)	0.55 (0.47,0.58)

AUC: area under the receiver-operating curve, LR: logistic regression, PTP: pre-test probability, PPV: positive predictive value, NPV: negative predictive value.

rately, and the performance metrics of the male-specific and female-specific model. 0.05 and 0.15 indicate the specific cut-off values at which an individual is classified as having CAD.

	Specificity	PPV/Precision	F1	NPV	AUC
	0.43 (0.42,0.44)	0.11 (0.11,0.12)	0.20 (0.20,0.21)	0.97 (0.97,0.98)	0.65 (0.63,0.66)
	0.81 (0.79,0.81)	0.18 (0.17,0.19)	0.27 (0.25,0.29)	0.95 (0.95,0.95)	0.65 (0.64,0.68)
	0.67 (0.65,0.68)	0.19 (0.18,0.19)	0.31 (0.30,0.32)	0.99 (0.98,0.99)	0.78 (0.77,0.79)
	0.91 (0.90,0.91)	0.36 (0.35,0.38)	0.45 (0.44,0.48)	0.96 (0.96,0.97)	0.76 (0.75,0.79)
	0.71 (0.69,0.72)	0.20 (0.19,0.21)	0.32 (0.31,0.33)	0.98 (0.98,0.99)	0.78 (0.77,0.79)
	0.91 (0.90,0.92)	0.37 (0.35,0.39)	0.45 (0.44,0.49)	0.96 (0.96,0.97)	0.75 (0.75,0.78)
	0.35 (0.34,0.37)	0.13 (0.12,0.14)	0.22 (0.22,0.24)	0.98 (0.97,0.99)	0.65 (0.63,0.65)
	0.60 (0.58,0.61)	0.16 (0.15,0.18)	0.27 (0.25,0.29)	0.96 (0.95,0.97)	0.68 (0.66,0.70)
	0.65 (0.63,0.67)	0.21 (0.20,0.23)	0.34 (0.33,0.37)	0.99 (0.98,0.99)	0.79 (0.77,0.80)
	0.89 (0.88,0.90)	0.39 (0.38,0.44)	0.50 (0.49,0.55)	0.97 (0.96,0.97)	0.80 (0.78,0.82)
	0.69 (0.67,0.71)	0.22 (0.22,0.25)	0.36 (0.35,0.40)	0.98 (0.98,0.99)	0.79 (0.78,0.81)
	0.89 (0.88,0.90)	0.37 (0.38,0.45)	0.48 (0.49,0.56)	0.96 (0.96,0.97)	0.78 (0.78,0.83)
	0.48 (0.47,0.50)	0.10 (0.09,0.10)	0.18 (0.16,0.18)	0.97 (0.96,0.97)	0.64 (0.61,0.65)
	0.97 (0.96,0.97)	0.33 (0.26,0.36)	0.27 (0.22,0.30)	0.94 (0.94,0.95)	0.60 (0.57,0.61)
	0.69 (0.65,0.70)	0.17 (0.14,0.17)	0.28 (0.24,0.28)	0.98 (0.98,0.99)	0.76 (0.75,0.78)
	0.92 (0.91,0.93)	0.33 (0.28,0.35)	0.40 (0.36,0.43)	0.96 (0.96,0.97)	0.71 (0.71,0.76)
	0.73 (0.70,0.74)	0.18 (0.15,0.18)	0.29 (0.26,0.30)	0.98 (0.98,0.99)	0.76 (0.74,0.78)
	0.94 (0.92,0.93)	0.36 (0.29,0.36)	0.42 (0.36,0.43)	0.96 (0.96,0.97)	0.72 (0.69,0.75)
	Specificity	PPV/Precision	F1	NPV	AUC
	0.35 (0.34,0.36)	0.13 (0.13,0.14)	0.23 (0.22,0.24)	0.98 (0.97,0.99)	0.64 (0.63,0.66)
	0.59 (0.58,0.61)	0.16 (0.16,0.17)	0.27 (0.26,0.29)	0.96 (0.95,0.97)	0.67 (0.66,0.70)
	0.62 (0.60,0.65)	0.20 (0.20,0.22)	0.33 (0.33,0.36)	0.99 (0.98,0.99)	0.77 (0.76,0.79)
	0.89 (0.87,0.90)	0.39 (0.37,0.42)	0.49 (0.48,0.53)	0.96 (0.96,0.97)	0.78 (0.77,0.82)
	0.69 (0.65,0.71)	0.23 (0.22,0.24)	0.37 (0.35,0.38)	0.98 (0.98,0.99)	0.79 (0.77,0.80)
	0.89 (0.88,0.90)	0.41 (0.38,0.43)	0.52 (0.49,0.54)	0.97 (0.96,0.97)	0.80 (0.78,0.82)
	Specificity	PPV/Precision	F1	NPV	AUC
	0.49 (0.47,0.50)	0.10 (0.09,0.10)	0.18 (0.16,0.18)	0.97 (0.96,0.97)	0.65 (0.61,0.65)
	0.97 (0.96,0.97)	0.35 (0.27,0.35)	0.28 (0.22,0.30)	0.95 (0.94,0.95)	0.60 (0.58,0.62)
	0.67 (0.63,0.68)	0.15 (0.14,0.16)	0.26 (0.24,0.27)	0.99 (0.98,0.99)	0.76 (0.74,0.77)
	0.93 (0.92,0.94)	0.33 (0.30,0.37)	0.40 (0.36,0.43)	0.96 (0.96,0.97)	0.71 (0.69,0.73)
	0.75 (0.69,0.75)	0.18 (0.16,0.19)	0.30 (0.27,0.30)	0.98 (0.98,0.99)	0.77 (0.76,0.79)
	0.93 (0.92,0.94)	0.34 (0.30,0.37)	0.42 (0.37,0.44)	0.97 (0.96,0.97)	0.74 (0.70,0.75)

Supplementary table 4 Baseline table of the distribution of patient characteristics and baseline variables between the train and test dataset.

	Overall	Train set	Test set	Missing (%)
n	34,524	25,893	8631	
Women (n, %)	19,141 (55.4)	14,340 (55.4)	4801 (55.6)	
Diagnosis of CAD (n, %)	2719 (7.9)	2039 (7.9)	680 (7.9)	
Age (mean (SD))	55 (14)	55 (14)	55 (14)	
BMI (mean (SD))	26.72 (4.81)	26.74 (4.81)	26.65 (4.83)	1.2
Classification of BMI (n, %)				1.2
Normal weight	13,147 (38.5)	9790 (38.3)	3357 (39.4)	
Obesity class I	5101 (15.0)	3809 (14.9)	1292 (15.2)	
Obesity class II	1458 (4.3)	1101 (4.3)	357 (4.2)	
Obesity class III	563 (1.7)	424 (1.7)	139 (1.6)	
Overweight	13,495 (39.6)	10,202 (39.9)	3293 (38.6)	
Underweight	355 (1.0)	267 (1.0)	88 (1.0)	
SBP (mean (SD))	141 (21)	141 (21)	140 (21)	1.4
DBP (mean (SD))	84.49 (11.78)	84.53 (11.75)	84.35 (11.88)	1.3
Having chest pain complaints (n, %)	28,466 (82.5)	21,313 (82.3)	7153 (82.9)	0
Having complaints of dyspnoea (n, %)	6661 (19.3)	5031 (19.4)	1630 (18.9)	0
Smoking status (n, %)				6.7
Current	13,032 (40.5)	9765 (40.5)	3267 (40.5)	
Ever	10,377 (32.2)	7781 (32.2)	2596 (32.2)	
Never	8803 (27.3)	6592 (27.3)	2211 (27.4)	
Diabetes (n, %)	2522 (7.3)	1923 (7.5)	599 (7.0)	0.5
Hypertension (n, %)	9514 (27.7)	7158 (27.8)	2356 (27.4)	0.4
Dyslipidaemia (n, %)	4911 (14.3)	3747 (14.5)	1164 (13.6)	0.5
Family history of atherosclerosis (n, %)				29.8
Negative	9523 (39.3)	7118 (39.1)	2405 (40.1)	
Unknown	2950 (12.2)	2233 (12.3)	717 (11.9)	
Positive	11,757 (48.5)	8875 (48.7)	2882 (48.0)	
Chest pain (n, %)				64.8
Non-anginal	7492 (61.7)	5594 (61.7)	1898 (61.6)	
Atypical	1704 (14.0)	1252 (13.8)	452 (14.7)	
Typical	2949 (24.3)	2220 (24.5)	729 (23.7)	
10-year SCORE CVD (median (IQR))	1.6 [0.4-4.7]	1.7 [0.4-4.7]	1.6 [0.4-4.5]	31.7
PTP (median (IQR))	6 [3-14]	6 [3-14]	6 [3-13.50]	
PTP category (n, %)				
<0.05	13,883 (40.2)	10,393 (40.1)	3490 (40.4)	
0.05-0.15	12,934 (37.5)	9683 (37.4)	3251 (37.7)	
>0.15	7707 (22.3)	5817 (22.5)	1890 (21.9)	

BMI: Body mass index, SBP: Systolic blood pressure, DBP: Diastolic blood pressure, SCORE CVD: Systematic COronary Risk Evaluation[34], PTP: pre-test probability, SD: standard deviation, IQR: interquartile range

Supplementary table 5 Overview of all features used for Lasso LR and their frequency of selection during bootstrapping in the general, male-specific and female-specific model.

Feature	General model	Female model	Male model
Patient number	500	500	500
Age	500	500	500
4-digit postal code	500	500	20
Missing dummy for 4th heart tone	500	79	498
Missing dummy for palpation of spleen	500	401	497
Missing dummy for patient's overall status	500	500	500
Systolic blood pressure	500	500	500
No complaints of chestpain	500	461	500
Complaints of dyspnoea	500	500	314
Any CVD risk factors at baseline	500	440	227
Stress ECG done during intake	500	452	497
Total cholesteral (lab)	500	435	495
Chest pain pressure	500	500	500
Radiation of chest pain	500	500	500
Duration of chest pain	500	500	500
Provocation of chest pain	500	500	500
Alleviation of chest pain	500	500	500
Number of present chest pain characteristics	500	497	500
Abnormal stress ECG	500	500	500
Incomplete stress ECG	500	159	500
Myocardial infarction during stress ECG	500	500	500
Myocardial infarction on ECG	500	372	500
Coronary dysfunction on ECG	500	500	500
Missing dummy for GFR (lab)	500	407	368
Missing dummy for troponin (lab)	500	500	500
Number of cigarettes per day	500	102	498
Family history of atherosclerosis	500	142	497

Feature	General model	Female model	Male model
Normal LV diastolic function (TTE)	500	500	498
Increased thickness of IMT (TTE)	500	153	500
Normal TTE	500	415	500
Antero-lateral ST-depression (X-ECG)	500	161	500
Infero-lateral ST-depression (X-ECG)	500	490	500
Typical chest-pain	500	500	500
Acceptable LV function (TTE)	499	348	453
Dyslipedemia	498	498	11
Pain in legs as reason to stop (X-ECG)	498	261	217
Number of alcohol units per day	498	414	353
Use of nitrate medication at baseline	493	142	375
E/e (TTE)	490	482	135
Load during stress ECG	489	16	500
Missing dummy for LV ejection fraction in biplane 4-chambers view (TTE)	488	92	446
Normal LV function (TTE)	488	320	496
Sinus rhythm during ECG	486	114	442
Heart rate during exercise (X-ECG)	483	494	142
Anterior ST-depression during exercise (X-ECG)	478	0	463
Antero-septal LV wall movement disorder (TTE)	473	5	488
Missing dummy for symptoms during exercise (X-ECG)	469	307	95
Normal ECG	465	11	434
Betablocker not stopped (X-ECG)	465	478	195
Missing dummy for systolic pulmonary artery pressure (TTE)	464	50	99
Triglycerides (lab)	464	172	235
Dyspnoea as reason to stop stress ECG	453	499	0

Feature	General model	Female model	Male model
Abnormal ECG	449	19	223
Missing dummy for lipoprotein-A (lab)	449	3	457
Haemoglobin (lab)	446	203	300
Changed ECG	436	152	339
GFR (lab)	434	22	393
Stenotic aortic valve (TTE)	430	1	218
Missing dummy for ventricular respons (TTE)	427	6	402
Velocity of LV outflow tract (TTE)	420	51	44
Atypical chest pain	416	0	458
Blood pressure as reason to stop stress ECG	412	6	440
P2Y12-inhibitor medication use at baseline	407	14	354
Moderate left atrial dilatation (TTE)	400	77	31
Typical onset of chest pain	388	125	53
Atrial fibrillation at ECG	385	348	86
LV hypertrophy at ECG	383	8	184
Diabetes	378	48	111
Chest pain as reason to stop stress ECG	376	8	399
Moderate concentric LV hypertrophy (TTE)	366	0	202
Missing dummy for auscultation	354	348	8
Blood pressure during stress ECG	352	342	77
Antero-lateral ST-depression (X-ECG)	331	94	123
Typical chest pain (classification before accounting for missings)	324	347	77
Postero-lateral ST-depression (X-ECG)	318	3	146
Unassessable RV dilatation (TTE)	317	108	128
Posterior LV wall movement disorder (TTE)	316	0	108
Missing dummy for glucose (lab)	311	0	5
Missing dummy for right atrial pressure (TTE)	304	13	173
PFO during valsalva (TTE)	302	36	57

Feature	General model	Female model	Male model
Valvular CVD risk factors at baseline	163	1	176
Missing dummy for bifurcation of right carotid artery (TTE)	159	57	11
LAB_TRIGclean.OneHot1	145	131	3
Metabolic equivalent measure of exercise intensity (X-ECG)	135	61	24
Infero-lateral ST-elevation (X-ECG)	135	92	11
Complaints of fatigue	134	30	0
Inferior ST-depression (X-ECG)	132	202	1
Inferior ST-elevation (X-ECG)	127	0	74
LAB_CHOL_LDLclean.OneHot1	126	6	23
Missing dummy for loss of decreased/changed conciousness during palpitation	123	3	0
Missing dummy for dorsalis pedis pulse	114	7	44
Missing dummy for free text about prodromes during collaps	112	0	165
Diameter of LV outflow tract (TTE)	111	0	2
Unknown use of alcohol	111	22	0
Decreased second heart sound	110	0	49
Mitral valve insufficiency IV (TTE)	108	0	8
No edema	107	1	9
Moderate right atrial dilatation (TTE)	105	1	244
No alcohol use	104	1	361
Antero-lateral pathological Q-wave (ECG)	103	0	0
Insulin use at baseline	101	76	2
Missing dummy for LV ejection fraction in biplane 2-chamber view (TTE)	100	44	48
LAB_CHOL_TOTclean.OneHot1	100	111	18
Missing dummy for presence of pain after pressure on thorax	94	24	0
Number of entry complaints	91	0	14

Missing dummy for progression of palpitations	300	18	2
Rhythm as reason to stop stress ECG	297	34	227
Poor LV function (TTE)	289	0	66
Assymetric/Obstructive LV hypertrophy (TTE)	279	4	0
No LV hypertrophy (ECG)	279	10	111
Use of metformin at baseline	276	67	47
Anterior ST-elevation (X-ECG)	274	0	198
Missing dummy for consult remarks	269	1	14
Mitral valve insufficiency III (TTE)	261	53	1
Inferior negative T-top (ECG)	256	161	13
Beta-blocker quitted NA (X-ECG)	250	5	422
Aortic valve velocity (TTE)	245	0	347
Use of aspirin at baseline	235	0	463
Target heart rate (X-ECG)	235	99	7
3/6 mitral valve insufficiency as first heart sound	235	0	62
Inferior pathological Q-wave (ECG)	228	51	66
Missing dummy for NT-proBNP (lab)	226	0	36
Missing dummy for thyroid stimulation hormone (lab)	222	0	183
Infero-lateral ST-depression (ECG)	219	28	103
Rhythm as indication for ECG	206	3	66
Missing dummy for bifurcation of left carotid artery (TTE)	205	4	80
LDL cholesterol (lab)	204	32	18
Normal RV function (TTE)	203	85	8
Mitral valve insufficiency I (TTE)	203	0	455
Tricuspid valve insufficiency II (TTE)	192	2	97
Borderline normal ECG	189	5	193
2/6 mitral valve insufficiency as first heart sound	185	0	224
Missing dummy for palpation of the liver	184	2	82
Unknown family history of sudden death	177	12	3
Septal LV wall motion disorder (TTE)	91	0	295
Missing dummy for sudden or gradial start of fatigue	87	0	2
Duration of stress ECG	87	10	29
Lateral LV wall motion disorder (TTE)	85	0	21
Missing dummy for frequency of dyspnoea	80	0	8
AV-NRT arrhythmia (X-ECG)	76	0	108
Creatinine (lab)	73	9	0
Posterior ST-elevation (ECG)	70	0	74
Normal kidney function based on GFR (lab)	68	62	1
Dyspnoea during chest pain	65	1	135
4/6 aortic stenosis as first heart sound	62	0	17
Missing dummy for hip circumference	61	11	0
No LV wall motion disorder (TTE)	61	1	44
Posterior ST-depression (X-ECG)	56	0	20
Use of NOAC at baseline	51	0	15
TTE done at intake	51	0	12
1/6 mitral valve insufficiency as first heart sound	50	13	0
Antero-septal ST-elevation (X-ECG)	50	19	7
Family history of sudden death	48	0	0
Anterior pathological Q-waves (X-ECG)	47	129	0
Moderate LV function (TTE)	44	4	6
Check-up for indication of ECG	43	0	19
Wide and fixed second heart sound	43	16	0
Run of premature ventricular complexes (X-ECG)	41	0	31
Antero-septal pathological Q-wave (ECG)	39	1	142
Anterior ST-depression (ECG)	39	0	265
Atrial fibrillation (X-ECG)	39	0	15
1/6 functional ejection as first heart murmur	38	1	13
Aortic valve insufficiency II (TTE)	38	1	0
Increased thickness of pulmonary valve (TTE)	37	219	0

Feature	General model	Female model	Male model
Obesity class I	33	0	41
Unknown family history of cardiomyopathy	32	4	0
Missing dummy for classification of dyspnoea	31	6	5
Missing dummy for QT duration (X-ECG)	31	0	115
Missing dummy for mean pressure gradient of mitral valve (TTE)	29	6	0
Inferior ST-depression (ECG)	29	213	27
Multiforme PVC (X-ECG)	29	0	83
Infero-posterior ST-depression (X-ECG)	29	0	4
Lateral ST-elevation (X-ECG)	29	5	174
Missing dummy for maximum pressure gradient of mitral valve (TTE)	28	1	2
ECG_CONCLUSION_AV1	27	0	36
LAB_K.OneHot1	27	0	29
Inferior LV wall motion disorder (TTE)	27	0	17
Infero-lateral LV wall motion disorder (TTE)	27	0	1
Tricuspid aortic valve (TTE)	27	1	3
Diastolic blood pressure	26	0	62
Complaints as indication for ECG	26	0	199
Lateral negative T-wave (ECG)	26	143	1
2/6 tricuspid valve as first heart murmur	25	0	0
Antero-septal ST-evalation (ECG)	25	0	5
No LV hyertrophy (X-ECG)	24	143	0
Supraventricular tachycardia (X-ECG)	24	1	0
Missing dummy for vena cava inferior diameter (TTE)	23	2	27
Anterior ST-elevation (ECG)	23	0	0
Anterior LV wall motion disorder (TTE)	21	35	0
Normal aortic flow (TTE)	21	0	0
Lateral ST-depression (ECG)	20	0	0
Anterior negative T-top (ECG)	10	3	77
Beta-blocker use at baseline	9	0	64
Glucose (lab)	9	6	0
Sinus rhythm (TTE)	9	1	82
Antero-lateral LV wall motion disorder (TTE)	9	0	8
Increased thickness of tricuspid valve (TTE)	9	0	6
Missing dummy for progression of fatigue	8	3	0
Missing dummy for frequence of palpations	8	0	0
Missing dummy for femoral pulse	8	25	0
Calcium-channel blocker use at baseline	8	0	59
Left atrial diameter (TTE)	8	0	47
HDL cholesterol (lab)	8	0	1
Fatigue as reason to stop stress ECG	8	18	0
Missing dummy for aspecific conduction delay (ECG)	8	0	1
Missing dummy for RBBB (ECG)	8	0	1
Atypical chest pain (classification before accounting for missings)	8	0	1
Side of blood pressure measurement	7	0	26
Missing dummy for LAHB (ECG)	7	0	1
Unknown family history of atherosclerosis	7	0	0
>50% collaps of inferior vena cava after inspiration (TTE)	7	0	17
Inferior ST-elevation (ECG)	7	0	6
Antero-septal ST-depression (X-ECG)	7	0	7
Missing dummy for Left ventricular posterior wall dimension at end-systole (TTE)	6	0	0
Missing dummy for LPHB (ECG)	6	0	0
Hypotensive blood pressure response (X-ECG)	6	0	5
Normal heart rate response (X-ECG)	6	0	51

Normal blood pressure response (X-ECG)	20	1	138
Alpha-blocker use at baselin	19	0	7
Slight aortic valve insufficiency (TTE)	19	0	8
Infero-lateral negative T-wave (ECG)	19	0	4
Conclusion of rhythm disorders (X-ECG)	18	1	0
Ventricular tachycardia (X-ECG)	18	101	0
Missing dummy for third heart sound	17	14	0
Antero-lateral negative T-top (ECG)	17	0	0
Missing dummy for diuresis	16	50	0
Antero-septal ST-elevation during rest (X-ECG)	16	24	1
E/A ratio (TTE)	15	0	22
Heart rate during rest (X-ECG)	15	0	6
Size of right posterior carotid artery (TTE)	14	0	0
ECG_CONCLUSION_IV1	14	0	9
No right atrial dilatation (ECG)	14	0	3
Inferior ST-depression during rest (X-ECG)	14	0	8
No arrhythmia (X-ECG)	14	6	0
Mildly decreased kidney function based on GFR (lab)	14	0	2
Missing dummy for triglycerides (lab)	13	0	0
Infero-posterior pathological q-wave (ECG)	13	0	5
Inferior pathological q-wave during rest (X-ECG)	13	213	19
Missing dummy for interventricular septum thickness at end-systole (TTE)	12	3	0
PR-interval (ECG)	12	0	1
Missing dummy for total cholesterol (lab)	12	0	0
Locally decreased breathing sounds	11	0	0
Slight right atrial dilatation (TTE)	11	0	18
Lateral ST-depression during rest (X-ECG)	11	0	17
Current smoker	10	0	45
Pulmonary valve insufficiency I (TTE)	10	15	0
Interventricular septum thickness at end-diastole (TTE)	5	0	1
Missing dummy for chest pain NYHA class	5	0	2
Missing dummy for LBBB (ECG)	5	0	0
LAB_GLUC.OneHot1	5	3	0
4/6 mitral valve as first heart sound	5	0	131
Dilated LV (TTE)	5	0	10
End-diastolic return over aortic valve (TTE)	5	13	0
No left atrial dilatation (ECG)	5	0	40
Lateral ST-segment during rest (X-ECG)	5	0	32
Premature atrial complexes (X-ECG)	5	9	0
Use of potassium-sparing diuretics at baseline	4	0	0
Missing dummy for descending aorta (TTE)	4	0	0
Maximum pressure gradient over aortic valve (TTE)	4	0	3
Missing dummy for pressure half-time (TTE)	4	1	0
Missing dummy for mitral valve area (TTE)	4	1	0
Screening as indication for ECG	4	0	0
Absence of second heart sound	4	0	36
Normal second heart sound	4	0	7
Underweight	4	0	0
Slight tricuspid valve insufficiency (TTE)	4	3	1
Antero-septal ST-depression (ECG)	4	0	34
Severely decreased kidney function based on GFR (lab)	4	0	4
Use of ezetimibe at baseline	3	0	10
Missing dummy for mean pressure gradient of aortic valve (TTE)	3	2	2
Corrected QT-interval (ECG)	3	72	0
Souffle in left and right carotid artery	3	0	0
2/6 functional ejection first heart sound	3	0	7
Locally increased breathing sounds	3	0	2

Feature	General model	Female model	Male model
Obesity class II	3	0	111
Positive family history for arrythmias	3	0	8
Unassessable RV function (TTE)	3	0	1
Slight dilation of left atrium (TTE)	3	1	1
Moderately decreased kidney function based on GFR (lab)1	3	2	0
Missing dummy for injury during collaps	2	0	0
Missing dummy for ascending aorta size (TTE)	2	1	0
Triscupid annular plane systolic excursion - movement lateral annulus (TTE)	2	0	0
Duration of RR interval (ECG)	2	0	1
QRS axis (ECG)	2	3	0
Lab measurements done at intake	2	0	0
Hypertension as indication for ECG	2	0	1
Normal central venous pressure	2	0	2
Non-dilated LV (TTE)	2	0	2
Prolaps of mitral valve (TTE)	2	0	6
Tricuspid valve insufficiency III (TTE)	2	0	0
Restrictive LV dysfunction (TTE)	2	0	1
Atrial flutter (ECG)	2	0	5
Antero-septal negative t-wave (ECG)	2	0	11
Sinus rhythm during rest (X-ECG)	2	0	3
Run of premature atrial complexes (X-ECG)	2	0	0
Uniform ventricular extrasystoles (X-ECG)	2	0	0
Missing dummy for ictus cordis	1	0	0
Missing dummy for auscultation of femoral artery	1	0	2
Missing dummy for symptoms of collaps	1	0	1
Weight	1	0	0
Complaints of palpitations	1	0	28
Missing dummy for type of collaps	0	3	2
Missing dummy for amount of collapses	0	0	12
Height	0	0	4
Hypertension	0	1	0
Arrhythmia risk factors at baseline	0	11	0
Use of loop diuretics at baseline	0	0	42
Use of statins at baseline	0	15	1
Left ventricular posterior wall dimension at end-diastole (TTE)	0	0	2
Fractional shortening (TTE)	0	0	9
Missing dummy for maximum dynamic pressure gradient over aortic valve (TTE)	0	0	24
Missing dummy for aortic valve area (TTE)	0	5	3
Missing dummy for left anterior carotis (TTE)	0	0	36
Missing dummy for right anterior carotis (TTE)	0	1	2
Relative wall thickness (TTE)	0	0	1
Left atrial diameter indexed by body surface area (TTE)	0	0	38
QT interval (ECG)	0	5	0
Missing dummy for selection as a candidate for heart revalidation (X-ECG)	0	2	0
Systolic blood pressure in rest (X-ECG)	0	0	2
Diastolic blood pressure (X-ECG)	0	0	1
Systolic blood pressure (X-ECG)	0	1	0
Vegetative symptoms of chest pain	0	0	1
Cardiovascular screening as indication for ECG	0	0	1
Abnormal ECG as indication for ECG	0	0	2
Second opinion as indication for ECG	0	0	2
LAB_HB.OneHot1	0	0	2
LAB_CREATclean.OneHot1	0	1	36

Use of ACEI at baseline	1	0	0
Use of ARBs at baseline	1	0	5
Use of thiazide diuretics at baseline	1	0	1
Use of sulphonylureas	1	0	1
Use of vitamin K antagonists at baseline	1	0	1
Dimension of aortic root (TTE)	1	0	2
Heart rate (ECG)	1	0	0
Missing dummy for RR-interval in rest (X-ECG)	1	0	4
Missing dummy for BNP (lab)	1	0	0
3/6 aortic stenosis murmur as first heart sound	1	0	2
Mid-systolic click as first heart sound	1	0	0
Unknown family history of arrhythmia	1	0	5
Concentric RV hypertrophy (TTE)	1	4	0
Severe left atrial dilatation (TTE)	1	0	104
Increased thickness of aortic valve (TTE)	1	0	6
Aortic valve insufficiency I (TTE)	1	0	10
Systolic return of pulmonary venous flow (TTE)	1	0	7
Pseudonormalization of LV diastolic function (TTE)	1	7	0
Antero-lateral ST-depression (ECG)	1	0	0
Atrial flutter in rest (X-ECG)	1	0	0
Lateral ST-depression (X-ECG)	1	4	6
Missing dummy of fatigue during exercise	0	0	3
Missing dummy of classification of fatigue	0	0	1
Missing dummy for type of palpitations	0	0	98
Missing dummy for speed of palpitations	0	0	6
Missing dummy for sudden/gradual start of palpitations	0	0	55
Missing dummy for thyroid gland	0	1	0
Increased central venous pressure during inspiration	0	1	3
Increased central venous pressure	0	0	1
1/6 aortic valve stensosis murmur as first heart sound	0	0	2
2/6 aortic valve insufficiency murmur as first heart sound	0	3	0
2/6 aortic valve stenosis murmur as first heart sound	0	0	11
Systolic murmur as first heart sound	0	1	0
Obesity class III	0	0	6
50 collaps of vena cava inferior after inspiration (TTE)	0	0	3
Assymetric LV hypertrophy (TTE)	0	1	0
Postero-lateral LV wall motion disorder (TTE)	0	2	0
Aneurysmatic/bulging RV dilatation (TTE)	0	1	0
Unassessable RV hypertrophy (TTE)	0	17	0
Acceptable RV function (TTE)	0	1	1
Severe right atrial dilatation (TTE)	0	0	8
Systolic anterior motion of mitral valve (TTE)	0	0	1
Increased thickness of mitral valve (TTE)	0	1	0
Tricuspid valve insufficiency I (TTE)	0	0	1
Aortic valve insufficiency III (TTE)	0	10	0
Slight pulmonary valve insufficieny (TTE)	0	4	0
Normal pulmonary venous flow (TTE)	0	0	1
Anterior pathological Q-wave (ECG)	0	0	15
Lateral pathological Q-wave (ECG)	0	0	1
Multiform ventricular extrasystoles (X-ECG)	0	6	0
Positive cardiovascular family history	0	1	0

X-ECG: stress ECG, TTE: transthoracic echocardiography, echography of carotid arteries, CVD: cardiovascular disease, GFR: glomerular filtration rate, LV: left ventricle, RV: right ventricle, PFO: patent foramen ovale, AV-NRT: AV-nodal re-entry tachycardia, RBBB: right bundle branch block, LAHB: left anterior hemiblock, NYHA: New York Heart Association.

Chapter 9

Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk

Klaske R. Siegersma[¥], Rutger R. van de Leur[¥], N. Charlotte Onland-Moret, David A. Leon, Ernest Diez-Benavente, Liesbeth Rozendaal, Michiel L. Bots, Ruben Coronel, Yolande Appelman, Leonard Hofstra, Pim van der Harst, Pieter A. Doevendans, Rutger J. Hassink, Hester M. den Ruijter[*], René van Es[*]

[¥,*] These authors contributed equally

EHJ - Digital Health 2022; 3(2): 245-254

ABSTRACT

Background Sex-differences in electrocardiograms (ECGs) are well documented and deep neural networks (DNN) based on the ECG accurately predict sex. Sex-differences in DNN models may lead to new insights on electrophysiological mechanisms and prognosis. Therefore, we validated DNN-based sex classification on ECGs and focused on visualizing features important for this classification. In addition, we analysed misclassification of sex and mortality risks.

Methods A DNN was trained to classify sex based on 131,673 normal ECGs. The algorithm was validated on an internal (68,500 ECGs) and external datasets (3,303 and 4,457 ECGs), respectively. The survival of sex (mis)classified groups was investigated using time-to-event analysis, and a sex-stratified mediation analysis was performed.

Results The DNN successfully distinguished females from males ECGs (internal validation: AUC 0.96 [95% CI 0.96-0.97]; external validations: AUC 0.89 [95% CI 0.88-0.90] and 0.94 [95% CI 0.93-0.94]). Sex-misclassified individuals (11%) had a 1.38 times higher mortality risk compared to correctly classified peers. Ventricular rate and QTc interval were mediating mortality risk in males, whilst in females, QRS was the strongest negative mediating factor. Indeed, a short QRS duration increased mortality risk in both sexes.

Discussion DNNs accurately classify sex based on raw ECG signals. While the proportion of misclassified individuals is low, a worse survival is seen in both sexes. This worse survival is mostly explained by known ECG features in misclassified males, but less so in females. By focussing on sex in DNNs, we uncovered a previously unknown ECG feature important for mortality.

INTRODUCTION

Despite increasing awareness on sex differences in cardiology, women remain underrepresented in randomized clinical trials.[1,2] Often, even when enough women are included, a focus on sex stratification is absent, despite pronounced differences between the sexes.[2,3] This also applies to the recent advancement of artificial intelligence (AI) in cardiology.[1] Deep neural networks (DNN) are increasingly used to analyse raw electrocardiogram (ECG) signals for prediction, diagnosis and prognosis of cardiovascular disease (CVD).[4–7] Sex differences in the ECGs are well known, as women have a higher heart rate, shorter PR and QRS duration, longer corrected QT duration (QTc), different T-wave morphology and lower precordial QRS and T-wave amplitudes than men.[8–10]

On top of applications to improve patient outcome in cardiovascular disease, DNNs are also able to classify sex based on ECG with an extremely high accuracy.[11] This suggests that the "black box" of DNNs may hold additional information on sex differences within ECGs that are currently unknown. Identifying subtle sex differences on the ECG will not only improve our understanding of these sex differences but might also be clinically relevant. ECG features that have already shown to be associated with mortality in a sex-dependent way are QRS prolongation, QTc prolongation, and T-wave morphology.[12–15]

As incorporating sex-stratified analysis into experimental design has enabled advancements across many disciplines, we hypothesize that a focus on sex differences using DNNs for analysing raw ECG signals could lead to new discoveries and insights.[1] For that purpose, we used two large data sources of normal ECGs to validate the high accuracy of the classification of sex with DNNs. In addition, we identified individuals who were misclassified on sex, and studied their survival. We highlight how these sex-specific ECG features affect mortality using visualizations techniques and mediation analyses.

METHODS

Study participants and data acquisition

UMCU TRAINING AND INTERNAL VALIDATION DATASET

All 10-second 12-lead resting ECGs (n=1,136,113) acquired in the University Medical Center Utrecht (UMCU) between July 1991 and December 2019 from individuals (n=249,262) aged between 18 and 85 years were selected. Demographic (age, sex, follow-up) and ECG data were extracted from hospital files of these individuals. All individuals (n=137,000) with at least one normal ECG (n=287,547) were selected (Figure 1). Only ECGs that were deemed interpretable were included.

The ECGs were recorded using a General Electric MAC V, 5000 or 5500 (GE Healthcare, Chicago, IL, USA) at 250 or 500 Hz and extracted in raw voltage format. Linear interpolation was used to resample every ECG to 500 Hz. The representative median beat was used

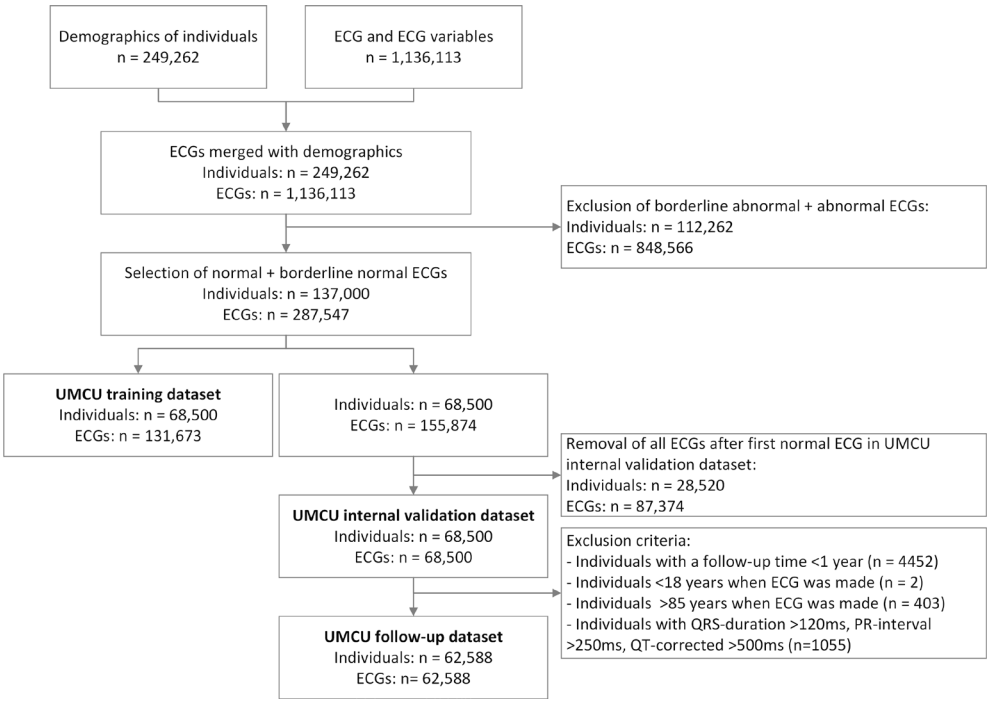


Figure 1 Selection of individuals for the UMCU training dataset, internal validation dataset and dataset for time-to-event and mediation analysis.

in this study and derived from these 10-second recordings by aligning all QRS complexes and taking the median voltage. R peaks were detected using the Stationary Wavelet Transform detector.16 Extraction of the conventional ECG features, such as PR interval, is described in more detail in the Supplementary Methods. All recordings obtained at non-cardiology departments were systematically annotated by a physician as part of the regular clinical workflow. The other ECGs were annotated by the Marquette 12SL algorithm (GE Healthcare, Chicago, IL, USA). Diagnostic ECG statements were extracted from these free text annotations using a text mining algorithm described before and used to determine if an ECG was interpreted as normal or borderline normal.[4]

UMCU FOLLOW-UP DATASET

A subset of the UMCU internal validation dataset was used to determine the association between ECG-classified sex and all-cause mortality. Survival data from all individuals were extracted from the Dutch Population Register. For these analyses individuals with less than one year of follow-up (n=4,452) and individuals with ECG conduction intervals outside normal ranges (QRS<120 ms, PR interval<250 ms and Bazett QTc<500ms; n=1,055) were excluded. This enabled investigation of long-term follow-up and avoids the bias that occurs because an individual is already in the hospital for a specific reason

that reduces the life expectancy (e.g. severe trauma or palliative care). These exclusions resulted in a final dataset of 62,588 individuals (figure 1).

EXTERNAL VALIDATION: KNOW-YOUR-HEART DATASET AND UTRECHT HEALTH PROJECT DATASET

External validation of the algorithm was performed in two external datasets. The Know-Your-Heart (KYH) dataset is a cross-sectional population-based study from two Russian cities, Arkhangelsk and Novosibirsk. This cohort consisted of 4647 participants, aged between 35 and 69 years. The full protocol of the KYH study has been described elsewhere.[17] A detailed description of the ECG acquisition in this dataset is provided in the Supplementary Methods.

The Utrecht Health Project (UHP) is an ongoing dynamic population study initiated in a newly developed large residential area in Leidsche Rijn, part of the city of Utrecht.[18] All new inhabitants were invited by their general practitioner to participate in the UHP. Written informed consent was obtained and an individual health profile (IHP) was made by dedicated research nurses. Survival data was obtained through the general practitioner via the International Classification of Primary Care (ICPC)-codes. The UHP study was approved by the Medical Ethical Committee of the University Medical Center, Utrecht, The Netherlands. UHP included baseline normal ECGs of 4457 individuals (2469 females, 55.4%), with a median age of 35 years [IQR 30-43]. The full protocol of the UHP cohort has been described elsewhere.[19]

Deep neural network development

A convolutional DNN architecture with several 1-dimensional causal dilated convolutional layers was trained to classify sex on ECG. This network architecture is inspired by van den Oord et al. and was described in detail previously.[20,21] The architecture had been previously optimized for use on median beats before and no further hyperparameter tuning was performed on this dataset.[21,22] Training was performed with a binary cross-entropy loss and the Adam optimizer with a learning rate of 0.0001 and batch size of 128.[23,24] Early stopping was performed when the validation did not decrease for 20 epochs. Output of the DNN was a probability that indicates the likelihood of an ECG belonging to a female individual. Cut-off value was set to 0.5, i.e. a probability <0.5 resulted in the classification of the ECG belonging to a male. All algorithm development was performed with the PyTorch package (version 1.7.0).[25]

Algorithm visualization

To determine what segments of the ECG are important for the DNN to classify sex, we used Guided Gradient Class Activation Mapping++ (Guided Grad-CAM++). This technique combines Grad-CAM, which provides global class-discriminative ECG segments, with guided backpropagation to achieve fine-grained timepoint-specific visualiza-

tions.[26,27] A detailed description on the visualization technique can be found in the Supplementary Methods.

Statistical analysis

DESCRIPTIVE STATISTICS OF DATASET AND PERFORMANCE EVALUATION OF DNN

The baseline characteristics of the datasets were described as mean +/- standard deviation (SD) or median with interquartile range (IQR), where appropriate. The discriminatory performance of the DNN in the UMCU internal validation set and KYH and UHP external validation set was assessed with the area under the receiver operating characteristic (AUC) and accuracy, calculated as the number of correctly classified individuals divided by the total number of individuals. The 95% confidence intervals (CI) around the performance measures were obtained using 2000 bootstrap samples. Four groups were identified for subsequent analyses using a predicted probability cut-off of 0.5: correctly classified males and females, biological females classified as male and biological males classified as female. Conventional ECG features (e.g. PR interval) were compared between these groups. No p-values were provided in these comparisons.

SURVIVAL AND MEDIATION ANALYSIS IN UMCU FOLLOW-UP DATASET

Using data from the UMCU follow-up and the UHP external validation dataset, sex-stratified survival analysis with Kaplan-Meier curves and Cox regression was done to evaluate the differences in survival between the four groups. All analyses were performed with age as the primary time variable (e.g. correction for late entry or left-truncation), as included individuals had their first ECG at different ages. Subsequently, the UMCU follow-up dataset wat used to investigate to what extent the relationship between sex (mis)classification by the DNN and survival is mediated by the conventional ECG features. Therefore, a biological sex-stratified mediation analysis was performed. In the mediation analyses, survival was modelled using an age-adjusted Weibull model and all ECG features were normalized. Mediation analysis was done for each ECG variable separately using the R mediation package (version 4.5.0).[28] The proportion effect explained PEE (i.e. how much of the effect of DNN-predicted sex on mortality is mediated by a conventional ECG variable) was derived by dividing the average causal mediated effect (ACME) by the total effect. We derived 95% CI around the PEE using nonparametric bootstrap with 1000 samples. Finally, a sex-stratified post-hoc evaluation of the association between conventional ECG features and all-cause mortality was performed to investigate non-linearities in the UMCU follow-up dataset. Therefore, all conventional ECG features were added to a Cox regression model using a natural cubic spline. Hazard ratios (HR) relative to the median value of the ECG variable were used to visualize the non-linear relationship.

All statistical analyses were executed using R version 3.5 (R Foundation for Statistical Computing). The Transparent Reporting of a Multivariable Prediction Model for Individu-

Table 1 Baseline table of the train and internal validation set from the UMCU database, stratified by sex.

	Overall	Train		Test	
		Males	Females	Males	Females
Individuals, n	137,000	34,214	34,286	34,617	33,883
ECGs, n	200,173	67,634	64,039	34,617	33,883
Deceased (n, %)	17,764 (16.9)	3415 (18.7)	2760 (15.3)	6495 (18.8)	5094 (15.0)
Age at ECG in years (median [IQR])	57 [44-68]	57 [46-67]	57 [43-68]	58 [46-68]	57 [43-68]
Linkage for follow-up possible (n, %)	104,848 (76.5)	18,283 (53.4)	18,065 (52.7)	34,617 (100)	33,883 (100)
Time between ECG and follow-up in years (median [IQR])	7.8 [3.5-13.8]	7.3 [3.2-12.9]	7.7 [3.5-13.8]	8.1 [3.5-14]	8.4 [3.8-14.4]
Age at ECG in years (n, %)					
≤ 30	19,426 (9.7)	6012 (8.9)	6583 (10.3)	3202 (9.2)	3629 (10.7)
31-40	20,576 (10.3)	6310 (9.3)	7285 (11.4)	3205 (9.3)	3776 (11.1)
41-50	32,920 (16.4)	11,354 (16.8)	10,989 (17.2)	5328 (15.4)	5249 (15.5)
51-60	44,114 (22.0)	16,310 (24.1)	13,115 (20.5)	7827 (22.6)	6862 (20.3)
61-70	47,875 (23.9)	17,052 (25.2)	13,970 (21.8)	9141 (26.4)	7712 (22.8)
71-80	29,045 (14.5)	9064 (25.2)	8913 (15.3)	5017 (14.5)	5151 (15.2)
> 80	6217 (3.1)	1532 (2.3)	2284 (3.6)	897 (2.6)	1504 (4.4)
Year of ECG (n, %)					
≤ 1995	32,167 (16.1)	15,132 (22.4)	14,876 (23.2)	1095 (3.2)	1073 (3.2)
1996-2000	34,249 (17.1)	14,447 (21.4)	13,984 (21.8)	3038 (8.8)	2780 (8.2)
2001-2005	22,648 (11.3)	5380 (8.0)	5452 (8.5)	5869 (17.0)	5947 (17.6)
2006-2010	31,995 (16.0)	8997 (13.3)	8709 (13.6)	7217 (20.8)	7072 (20.9
2011-2015	44,709 (22.3)	13,547 (20.0)	11,713 (18.3)	9815 (28.4)	9634 (28.4)
> 2015	34,405 (17.2)	10,131 (15.0)	9314 (14.5)	7583 (21.9)	7377 (21.8)
Ventricular rate (median [IQR])	71 [62-82]	69 [60-81]	72 [64-83]	69 [60-80]	72 [63-82]
PR interval, ms (median [IQR])	154 [140-170]	156 [142-174]	150 [136-168]	158 [142-174]	150 [136-166]
QRS duration, ms (median [IQR])	90 [84-98]	96 [88-100]	86 [80-92]	96 [88-102]	86 [80-94]
QT interval, ms (median [IQR])	386 [364-408]	386 [364-410]	384 [360-406]	388 [364-410]	384 [364-408]
Bazett corrected QT interval, ms (median [IQR])	417 [404-434]	414 [402-430]	420 [407-436]	414 [401-430]	419 [406-437]
SL voltage (median [IQR])	1.94 [1.6-2.4]	2 [1.6-2.4]	1.91 [1.6-2.3]	2 [1.6-2.4]	1.9 [1.5-2.3]
Cornell voltage (median [IQR])	1.3 [0.9-1.6]	1.3 [1-1.7]	1.1 [0.8-1.5]	1.4 [1-1.8]	1.2 [0.9-1.5]
SL product (median [IQR])	175 [139-218]	187 [148-232]	165 [133-204]	187 [148-232]	165 [131-202]
Cornell product (median [IQR])	112 [81-150]	126 [93-164]	97 [69-128]	132 [97-171]	101 [73-133]

ECG: electrocardiogram, ms: milliseconds, IQR: interquartile range, SL: Sokolow-Lyon.

Table 2 Overview of baseline and ECG features in the UMCU follow-up dataset and the distribution

		Males	
	Overall	Correctly classified	
Individuals-n	31,328	26,344	
Deceased (n-%)	4066 (13.0)	3213 (12.2)	
Age at ECG in years (median [IQR])	57 [45-67]	57 [44-66]	
Time between ECG and follow-up in years (median [IQR])	8.7 [4.4-14.5]	9 [4.6-14.9]	
Age at ECG in years (n-%)			
≤ 30	3052 (9.7)	2699 (10.2)	
31-40	3048 (9.7)	2672 (10.1)	
41-50	5020 (16.0)	4333 (16.4)	
51-60	7198 (23.0)	6115 (23.2)	
61-70	8137 (26.0)	6744 (25.6)	
71-80	4287 (13.7)	3344 (12.7)	
> 80	586 (1.9)	437 (1.7)	
Year of ECG (n-%)			
≤ 1995	1076 (3.4)	946 (3.6)	
1996-2000	2872 (9.2)	2385 (9.1)	
2001-2005	5351 (17.1)	4701 (17.8)	
2006-2010	6453 (20.6)	5470 (20.8)	
2011-2015	8917 (28.5)	7408 (28.1)	
> 2015	6659 (21.3)	5434 (20.6)	
Ventricular rate-bpm (median [IQR])	68 [60-79]	67 [59-77]	
PR interval in ms (median [IQR])	158 [144-174]	158 [144-174]	
QRS duration in ms (median [IQR])	96 [88-102]	96 [90-102]	
QT interval in ms (median [IQR])	388 [366-410]	388 [366-410]	
Corrected QT in ms (median [IQR])	413 [401-429]	411 [400-427]	
SL-voltage (median [IQR])	2 [1.6-2.4]	2 [1.6-2.4]	
Cornell-voltage (median [IQR])	1.4 [1.1-1.8]	1.4 [1.1-1.8]	
SL-product (median [IQR])	188 [149-233]	193 [153-238]	
Cornell-product (median [IQR])	132 [98-171]	135 [101-174]	

ECG: electrocardiogram, bpm: beats per minute, ms: milliseconds, IQR: interquartile range, SL: Sokolow-Lyon.

al Prognosis or Diagnosis Statement for the reporting of diagnostic models was followed, where appropriate.[29]

RESULTS

Characteristics of study population

Median age of included individuals at the time of their ECG acquisition in the UMCU dataset was 57.2 [IQR 44.7-67.6] years. Table 1 shows the baseline characteristics of the ECGs used in the UMCU training and internal validation set separated for males and females

between correctly classified males and females and their misclassified biological peers.

		Females		
	Misclassified	Overall	Correctly Classified	Misclassified
	4984	31,260	30,008	1252
	853 (17.1)	3188 (10.2)	2984 (9.9)	204 (16.3)
	61 [49-70]	56 [42-67]	56 [42-67]	61 [49-72]
	7.5 [3.7-12.8]	8.9 [4.6-15.0]	8.9 [4.6-15.0]	10 [4.9-15.5]
	353 (7.1)	3546 (11.3)	3451 (11.5)	95 (7.6)
	376 (7.5)	3639 (11.6)	3546 (11.8)	93 (7.4)
	687 (13.8)	5009 (16.0)	4852 (16.2)	157 (12.5)
	1083 (21.7)	6400 (20.5)	6136 (20.4)	264 (21.1)
	1393 (27.9)	7054 (22.6)	6743 (22.5)	311 (24.8)
	943 (18.9)	4596 (14.7)	4334 (14.4)	262 (20.9)
	149 (3.0)	1016 (3.3)	946 (3.2)	70 (5.6)
	130 (2.6)	1065 (3.4)	1014 (3.4)	51 (4.1)
	487 (9.8)	2658 (8.5)	2569 (8.6)	89 (7.1)
	650 (13.0)	5570 (17.8)	5251 (17.5)	319 (25.5)
	983 (19.7)	6446 (20.6)	6166 (20.5)	280 (22.4)
	1509 (30.3)	8915 (28.5)	8619 (28.7)	296 (23.6)
	1225 (24.6)	6606 (21.1)	6389 (21.3)	217 (17.3)
	74 [64-86]	71 [63-81]	71 [63-81]	69 [61-80]
	154 [138-170]	150 [136-166]	150 [136-166]	158 [144-174]
	90 [84-96]	86 [80-94]	86 [80-92]	92 [86-100]
	382 [356-408]	386 [364-408]	386 [364-408]	388 [366-412]
	422 [407-439]	419 [406-436]	419 [406-436]	416.50 [405-435]
	1.8 [1.5-2.2]	1.9 [1.6-2.3]	1.9 [1.6-2.3]	2 [1.6-2.4]
	1.3 [1-1.7]	1.2 [0.9-1.5]	1.2 [0.9-1.5]	1.3 [1-1.7]
	163 [131,202]	166 [132-203]	165 [132-202]	182 [146-227]
	117 [86-153]	100 [72-132]	99 [72-131]	122 [89-159]

and the distributions of the ECG features. Follow-up was available for 104.848 (76.5%) of individuals and median follow-up time (Table 2) for the UMCU follow-up dataset was 8.7 [IQR 4.4-14.5] years and 8.9 [IQR 4.6-15.5] for, respectively, males and females. Table 3 shows the distributions of the ECG features, stratified by sex and sex-classification, for the KYH external validation dataset. The baseline characteristics of the UHP external validation dataset are displayed in Table 4, stratified by sex and sex-classification by DNN. Follow-up was available in this dataset (median follow-up: 16.9 years [IQR 15.3-18.0]), but ECG features were not structurally reported.

Table 3 Overview of baseline and ECG features in KYH and UHP external validation dataset and the distribution between correctly classified males and females and their misclassified biological peers. For KYH no follow-up information was available, while for UHP the PR interval was not available.

	Know-Your-Heart				Utrecht Health Project			
	Males		Females		Males		Females	
	Correctly classified	Misclassified	Correctly Classified	Misclassified	Correctly classified	Misclassified	Correctly Classified	Misclassified
Individuals, n	902	412	1787	202	1872	116	1764	705
Deceased (n, %)	-	-	-	-	40 (2.1)	2 (1.7)	17 (1.0)	11 (1.6)
Age at ECG in years (median [IQR])	53 [45-61]	56 [47-64]	54 [45-62]	57 [49-64]	36 [31-43]	40 [34-51]	34 [29-43]	33 [29-41]
Time between ECG and follow-up	-	-	-	-	16.9 [15.3-18.0]	15.9 [15.2-17.2]	16.7 [15.2-17.9]	17.4 [15.6-18.1]
Ventricular rate in bpm (median [IQR])	63 [57-70]	62 [56-70]	64 [59-71]	63 [58-69]	60 [54-66]	60 [54-72]	66 [60-72]	60 [54-72]
PR interval, ms (median [IQR])	158 [146-174]	156 [140-174]	152 [138-168]	154 [142-168]	-	-	-	-
QRS duration, ms (median [IQR])	96 [90-102]	94 [88-102]	90 [84-96]	92 [86-98]	98 [92-104]	94 [86-102]	86 [80-92]	88 [82-96]
QT interval, ms (median [IQR])	402 [384-422]	406 [386-426]	410 [392-430]	410 [392-428]	394 [376-414]	395 [370-425]	394 [376-412]	394 [376-414]
Corrected QT, ms (median [IQR])	411 [398-426]	416 [402-430]	425 [412-438]	421 [409-435]	396 [381-412]	408 [390-417]	406 [392-422]	405 [389-420]
SL-voltage, mV (median [IQR])	2.2 [1.8-2.6]	2 [1.6-2.4]	2 [1.6-2.3]	2 [1.6-2.4]	2.3 [1.9-2.8]	2.1 [1.6-2.5]	1.9 [1.6-2.3]	2.0 [1.6-2.4]
Cornell-voltage, mV (median [IQR])	1.5 [1.2-1.8]	1.3 [1-1.7]	1.2 [0.9-1.5]	1.5 [1.1-1.8]	1.1 [0.8-1.5]	1.1 [0.8-1.4]	0.8 [0.5-1.1]	0.8 [0.5-1.1]
SL-product, mV (median [IQR])	210 [168-248]	189 [154-229]	177 [145-212]	183 [147-227]	222 [180-274]	193 [155-239]	164 [136-197]	172 [139-208]
Cornell-product, mV (median [IQR])	142 [108-176]	128 [95-162]	105 [78-137]	141 [103-169]	111 [79-147]	105 [70-129]	64.8 [44-91.3]	69.8 [43.9-101.5]

ECG: electrocardiogram, bpm: beats per minute, ms: milliseconds, IQR: interquartile range, SL: Sokolow-Lyon.

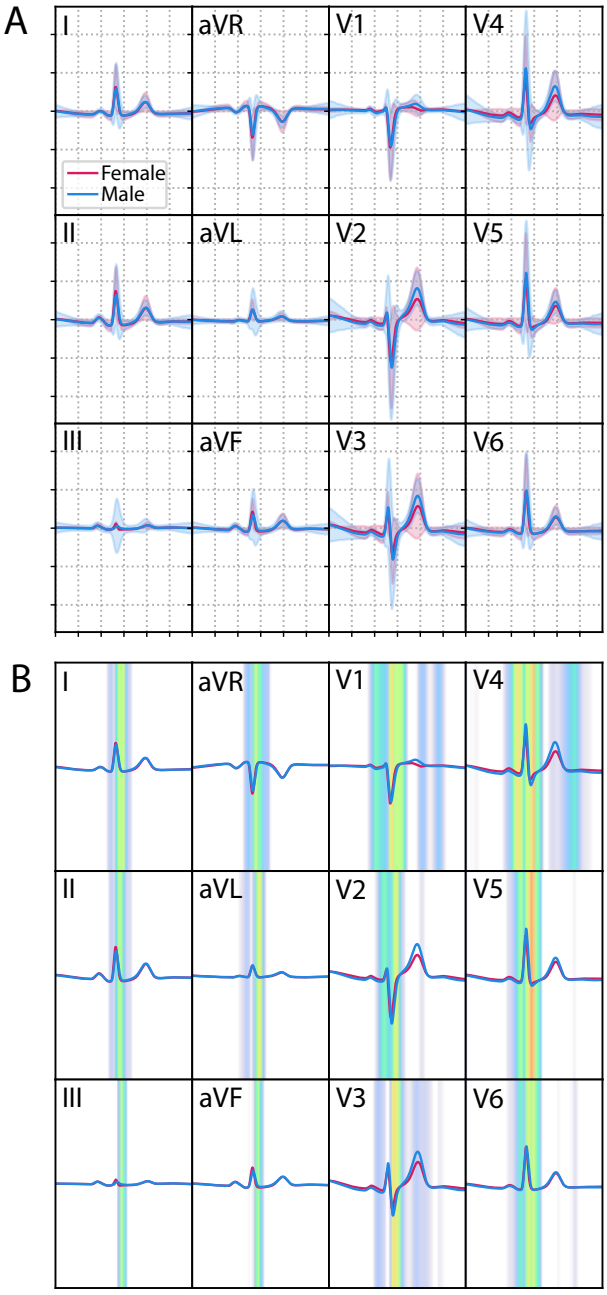


Figure 2 Guided Grad-CAM of the time-normalized median beats of the sex classification DNN. A: Overlay of median beats of biological males (blue) and females (red), showing where differences in the ECG occur. B: Guided Grad-CAM of the ECG, with highlighted areas of the ECG where the DNN focusses upon to give a classification.

Sex classification and feature detection

The AUC for sex classification with DNN in the UMCU internal validation dataset (n=68.500) was 0.96 (95% CI 0.96-0.97), with an accuracy of 0.89 (95% CI 0.89-0.89). The Guided Grad-CAM visualization algorithm showed that the DNN made its decisions mostly on the S-waves, especially in leads V3 to V5, and on the T-waves in V1 to V4 (Figure 2).

Evaluation of the individuals that were misclassified on sex in the UMCU follow-up dataset showed that they were older than their correctly classified biological peers. The median age of misclassified females was 61.3 [IQR 49.2-71.7] years versus 56.2 [IQR 42.1-67.3] years for correctly classified females, and 61.9 [IQR 49.0-69.8] years and 56.7 [IQR 44.2-66.2] years for, respectively, misclassified and correctly classified males. Overall, misclassified individuals had ECG characteristics resembling those of their biological counterparts (Table 2). Thus, males classified as females had higher ventricular rate, shorter PR and QRS duration, longer QTc and lower Sokolow-Lyon and Cornell voltages than correctly classified males. Females clas-

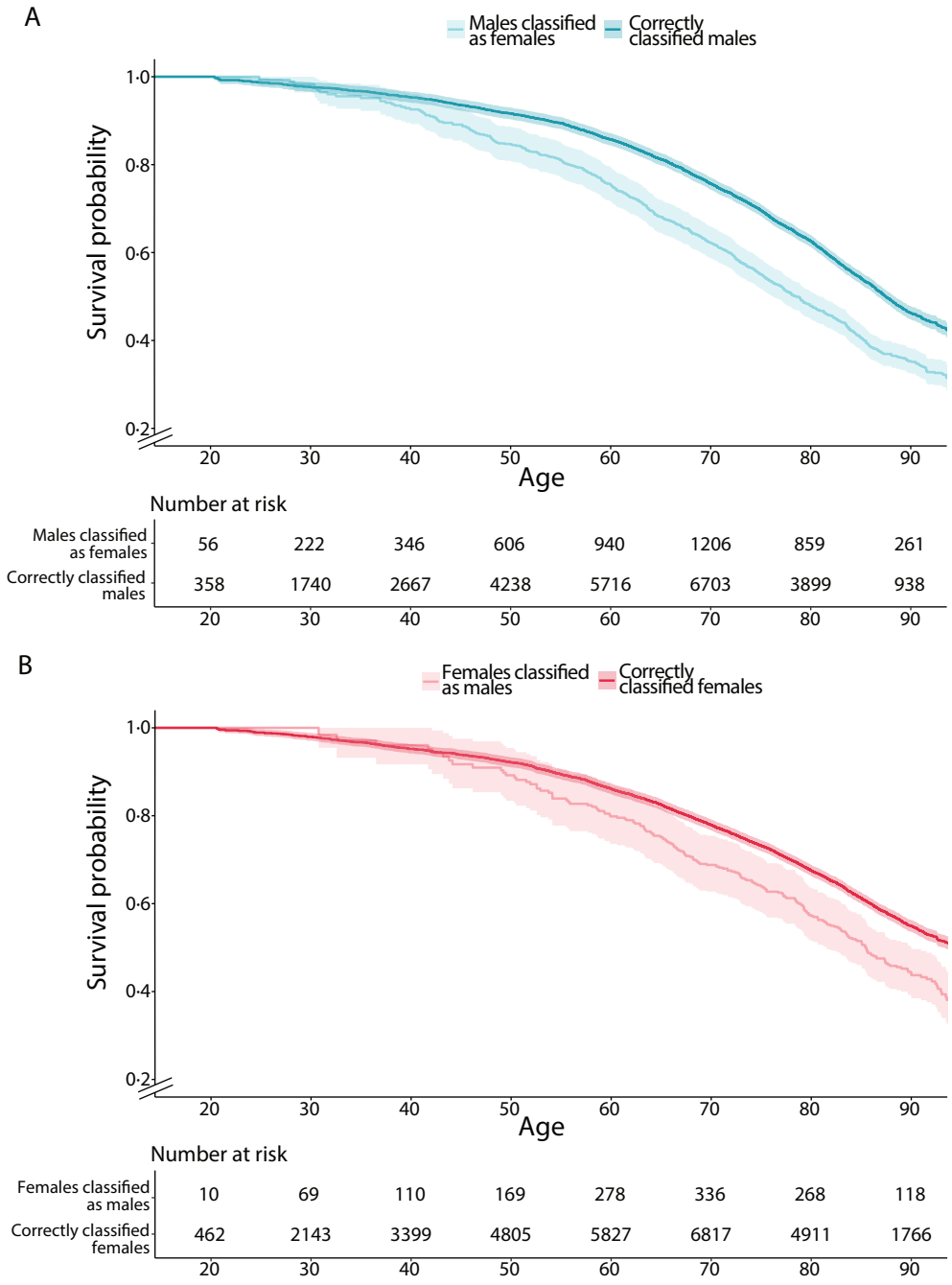


Figure 3 Kaplan-Meier curves (left-truncated) of males (A) and females (B), separately plotted for correctly classified and misclassified groups, corrected for late entry and stratified by classification of sex as was the output of the DNN.

sified as males, as compared to correctly classified females, had a longer PR and QRS duration and shorter QTc. An overview of all 83 features and their corresponding values within the (mis)classifications can be found in Supplementary table 1.

External validation using Know-Your-Heart and UHP dataset

The AUC and accuracy for the DNN in the KYH external validation dataset were, respectively, 0.89 (95% CI 0.88-0.90) and 0.81 (95% CI 0.80-0.82). Similar trends regarding the distribution of ECG features as in the UMCU internal validation dataset were seen when different classifications were compared (Table 3), e.g. misclassified individuals were overall older (median age misclassified females: 57.1, IQR 48.8-64.3, vs correctly classified females: 53.6, IQR 45.0-61.7, median age misclassified males: 56.0, IQR 47.3-63.7, vs correctly classified males: 53.2, IQR 44.7-61.1).

The AUC and accuracy for the DNN in the UHP dataset were, respectively, 0.94 (95% CI 0.93-0.94) and 0.82 (95% CI 0.80-0.83). In this dataset, the individuals were overall younger than in the KYH dataset and UMCU internal validation dataset with a median age of 36 [IQR: 31-44] for males and 34 [IQR: 29-42] for females. Yet, only misclassified males were older than their correctly classified biological peers (median age: 40, IQR: 34-50 vs 36, IQR: 31-43). The median age between misclassified and correctly classified females did not differ (misclassified females: 33, IQR: 29-41 vs correctly classified females: 34, IQR: 29-43).

Sex-specific survival analysis

In the UMCU follow-up dataset 3188 (10%) of included females and 4066 (13%) of included males died during follow-up. Mortality risk in this dataset was higher for biological males compared to biological females (HR: 1.33, 95% CI 1.27-1.39). In both sexes, a higher proportion of misclassified individuals died compared to their correctly classified biological peers: 16.3% (n=204) versus 9.9% (n=2984) of misclassified and correctly classified females, 17.1% (n=853) versus 12.2% (n=3213) of misclassified and correctly classified males. This was also shown in the Kaplan-Meier curves of both sexes (Figure 3) and confirmed by Cox regression with left-truncation that showed misclassified individuals had a higher mortality risk referenced to their correctly classified biological peers (HR misclassified females: 1.38, 95% CI 1.20-1.59, HR misclassified males: 1.38, 95% CI 1.28-1.49). Follow-up was also available for the individuals in the UHP external validation dataset. In this dataset mortality was low, with 28 (1.1%) females and 42 (2.1%) males who died. The mortality risk for males was higher compared to females (HR: 1.62, 95% CI 1.01-2.62). The increased mortality risk in misclassified individuals could only be confirmed, although not significant, for females in the UHP external validation dataset (HR misclassified females: 1.61, 95% CI 0.76-3.46 and HR misclassified males: 0.39, 95% CI 0.09-1.64).

Sex-stratified mediation analysis of ECG features and survival

Sex-stratified mediation analyses in the UMCU follow-up dataset showed that the relationship between misclassification of sex and mortality is mediated, at least in part, by conventional ECG features in both sexes. In females, the amplitude of the S-wave in lead V1 had a proportion effect explained (PEE) of 18% (95% CI 10%-35%, Figure 4). Also, the S-wave voltage in V4 (PEE 14%, 95% CI 7%-26%) and the ST-segment peak-to-peak voltage in lead I (PEE 14%, 95% CI 8%-25%) were mediators in the relation, although to a lesser extent. Multiple ECG features negatively mediated the relation in females, of which QRS duration to the largest extent (PEE -25%, 95% CI -52%- -15).

In males, ECG features were stronger mediators than in females, with the highest PEE of 39% (95%-CI: 31%-54%) for ventricular rate. Other mediators in males were QTc and the T-wave amplitude in V2, that mediated, respectively, 21% (95% CI 16%-29%) and 18% (95% CI 11%-28%). The beta coefficients and PEE of all normalized conventional ECG parameters in the Weibull model are shown in Supplementary tables 2 and 3.

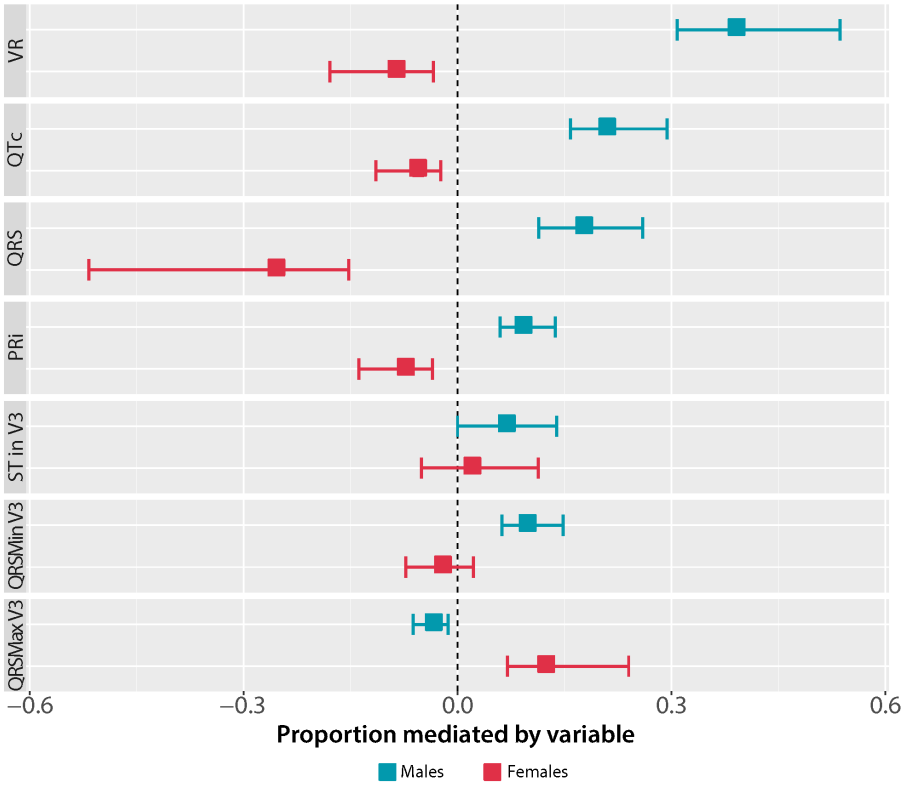


Figure 4 Proportion of the relation between classification and survival that is mediated by a selection of standard ECG features, stratified for females (red) and males (blue).
VR: Ventricular rate, QTc: Corrected QT according to Bazett, QRS: QRS duration, PRi: PR interval, ST in V3: maximum amplitude of T peak in V3, QRSMax V3: maximum amplitude of R peak in V3, QRSMin V3: minimum amplitude of S peak in V3.

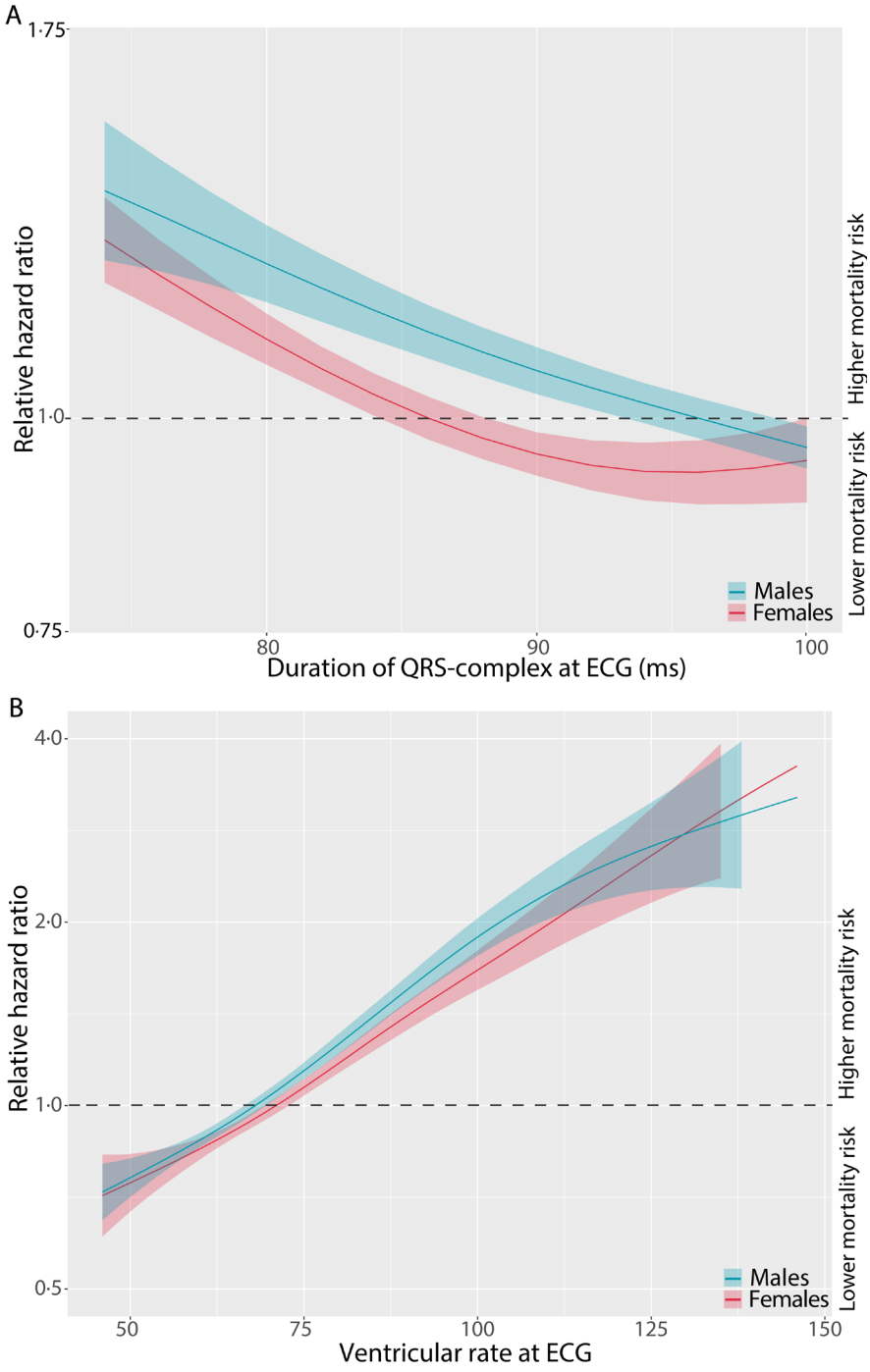


Figure 5 Relation between QRS duration (ms) (A) and ventricular rate (B) and hazard ratio in males and females.

Sex-specific analysis of QRS-duration, ventricular rate and mortality in all individuals

Two ECG features were highlighted (Figure 5) in the post-hoc analysis due to their large (negative) PEE in the mediation analysis. This analysis was performed in a sex-stratified manner in the UMCU follow-up dataset. First, QRS duration showed the start of a non-linear curve in the post-hoc analysis: individuals with short QRS duration have a higher mortality risk compared to individuals with a median QRS duration. HR increased again for increasing QRS duration. A higher mortality risk for individuals with an extended QRS duration was not observed in the graph since all individuals with a QRS duration >120ms were removed from the dataset. This relation between short QRS duration and higher mortality risk was confirmed in males in the UHP external validation dataset (Supplementary Figure 2). Second, ventricular rate showed an exponential relationship in both sexes with hazard ratio, displayed as a straight line on the logarithmic scale.

DISCUSSION

DNNs have excellent performance in classifying sex from ECGs, both in internal and external validation datasets. Misclassification of sex was associated with a higher mortality risk, independent of age. Subsequent mediation analyses showed that conventional ECG features mediate the association with mortality in misclassified males, but less so in females in whom the QRS interval showed an unexpected relation with mortality. A more elaborate analysis of this ECG feature showed an association with reduced survival in males and females, which has not previously been described. This finding was reproduced for males only in an external validation dataset. Our study highlights the importance of studying sex differences with AI to uncover new biology.

Our study shows a similar performance of the DNN on median beats for classification of sex as was described on full 10-second ECG.[11] Yet, the study by Attia et al.[11] stated the necessity to understand the relevance of discordance between ECG-classified and true biological sex for the individual. Our analysis of ECG features focused on this knowledge gap and showed that misclassification occurs when the ECGs become more alike, i.e. females that have ECG features similar to males are more often misclassified and the same holds for males that have ECG features similar to females, which was confirmed in external validation. Furthermore, discordance between ECG-classified and true biological sex was associated with a reduced survival in both sexes.

Our initial hypothesis was that the survival curve of misclassified individuals would move-towards their biological counterparts, i.e. survival of incorrectly classified males would be better than correctly classified males, while the survival of misclassified females would be worse than that of correctly classified females. However, our study shows that misclassification was associated with worse survival for both sexes in the UMCU follow-up dataset. Yet, in the UHP external validation dataset, HR fit the hypothesis (HR misclassified

females: 1.61, 95% CI 0.76-3.46 and HR misclassified males: 0.39, 95% CI 0.09-1.64), but were not statistically significant. Mediation analyses showed that increased mortality risk was largely mediated by ventricular rate and QTc in males, but conventional ECG features could not explain this phenomenon in females. This result indicates that the DNN is picks up subtle changes in the ECG that are not included in conventional ECG features and profoundly affect sex classification and survival.

Surprisingly, the non-linear post-hoc analysis showed increased mortality risk in the UMCU follow-up selection with a shortening QRS duration (<80ms), which was validated in the UHP external validation dataset. After an optimum at 100ms for females, a higher QRS duration is again associated with worse survival. As this study only included normal ECGs, we could not confirm the shape of the relation between QRS duration and HR above 120ms, but previous studies have shown higher HRs for increased QRS duration in both sexes.[14] It can therefore be assumed that for males the optimum of the J-curve is around 100ms. The nonlinear relationship between QRS duration and mortality has, to our knowledge, not been previously identified. However, it has been hypothesized that increasing extension of the Purkinje system into the walls of the ventricular system is associated with a shorter QRS duration, which make these individuals more at risk for idiopathic ventricular re-entrant arrhythmias.[30]

Other important mediators were increased ventricular rate and QTc. Increased ventricular rate in our study was associated with an increased HR for both sexes. This association has been previously shown in multiple studies.[31–35] It is assumed that a higher resting heart rate is a measure of an overall worse physical condition.[32,34] Our mediation analyses also showed a large PEE by QTc for the association between classification and survival in males. A meta-analysis of available literature showed increased association between mortality and a longer QTc, also when the QTc is within normal boundaries, confirmed in our post-hoc analysis (Supplemental figure 1).[12]

Strengths in this study are, first, the large amount of annotated ECG data that has been used, which gave the opportunity to only select normal ECGs. Second, Guided Grad-CAM was used to visualize important regions for the DNN to classify sex[26], enlightening the black box of DNN and ECG differences between the sexes. The Grad-CAM filter hinted toward specific features that are different between the sexes. Evaluation of the median values of the ECG in the different classifications confirmed these differences (Supplementary table 1). Third, this study was the first study to externally validate a sex classification algorithm.

Furthermore, our study included many females, which gave us the opportunity to specifically study sex differences and perform all analyses in a sex-stratified manner. As women remain underrepresented in clinical research in the cardiovascular domain, regular care data is crucial to address pressing cardiovascular topics in women.[36,37] Despite more

awareness, sex-stratification is often not performed, which also applies to validation of AI algorithms.[1] This study is unique in that it provided new insights into sex-specific ECG features that are associated with mortality, through the classification of sex with ECG-based AI. The presented results feed future research into sex-specific conductivity mechanisms that influence survival, unravelling the conundrum of sex-differences in longevity.

This study has some limitations. First, the hospital-visiting population that was used in this study had an ECG for a specific reason, although we selected only ECGs classified as 'normal' by either the ECG recording software or the examining physicians. Underlying pathophysiology that does not directly affect the ECG, but also stress and anxiety related to an out-patient clinic or hospital visit, might induce subtle changes on the ECG, including an increase in heart rate. This could make the DNN less generalizable to non-hospital populations. However, our external validation analyses showed the decrease in performance to be limited. Second, no causes of death were known for the UMCU follow-up dataset, which prevented us to look specifically into the analysis of cardiovascular mortality. Third, the UHP external validation dataset was significantly different from the UMCU internal validation dataset. In general, individuals in the UHP dataset were younger. This, in combination with a low number of events, might have caused our inability to show a significant reduced survival risk in misclassified individuals. Yet, we were able to validate the relation between shortening of the QRS duration in the UHP external validation dataset for males, which is a promising feature for future studies. Also, the UHP and KYH external validation datasets are cohort populations and thus inherently different from the hospital-visiting population used for development of the DNN. Nevertheless, performance of the DNN on both populations for classification of sex was still excellent. Fourth, survival information in the KYH external validation cohort was unavailable. Therefore, our findings regarding survival and the influence of different ECG features of misclassified individuals warrants further validation. However, the pattern of differences in the ECG characteristics according to classification status was replicated in the Know-Your-Heart dataset, which comprised a random population sample.

DNNs accurately classify sex based on raw ECG signals. While the proportion of misclassified individuals is low, a worse survival is seen in both sexes. This worse survival is mostly explained by known ECG features in misclassified males, but less so in females. Based on mediation analysis in females, we identified an unexpected relation between a short QRS complex and mortality. This study shows that focussing on sex differences in DNNs is useful to uncover previously unknown ECG features important for mortality.

REFERENCES

1. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. *Nature*. 2019;575(7781):137-146. doi:10.1038/s41586-019-1657-6
2. Bots SH, den Ruijter HM. Recommended heart failure medications and adverse drug reactions in women call for sex-specific data reporting. *Circulation*. 2019;139(12):1469-1471. doi:10.1161/CIRCULATIONAHA.118.037585
3. Vogel B, Acevedo M, Appelman Y, et al. The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030. *Lancet*. 2021;397(10292):2385-2438. doi:10.1016/S0140-6736(21)00684-X
4. van de Leur RR, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks. *J Am Heart Assoc*. 2020;9(10):e015138. doi:10.1161/JAHA.119.015138
5. Ko WY, Siontis KC, Attia ZI, et al. Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram. *J Am Coll Cardiol*. 2020;75(7):722-733. doi:10.1016/j.jacc.2019.12.030
6. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394(10201):861-867. doi:10.1016/S0140-6736(19)31721-0
7. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nat Med*. 2019;25(1):70-74. doi:10.1038/s41591-018-0240-2
8. Rijnbeek PR, Van Herpen G, Bots ML, et al. Normal values of the electrocardiogram for ages 16-90 years. *J Electrocardiol*. 2014;47(6):914-921. doi:10.1016/j.jelectrocard.2014.07.022
9. van der Ende MY, Siland JE, Snieder H, van der Harst P, Rienstra M. Population-based values and abnormalities of the electrocardiogram in the general Dutch population: The LifeLines Cohort Study. *Clin Cardiol*. 2017;40(10):865-872. doi:10.1002/clc.22737
10. Simonson E, Blackburn H, Puchner TC, Eisenberg P, Ribeiro F, Meja M. Sex Differences in the Electrocardiogram. *Circulation*. 1960;22(4):598-601. doi:10.1161/01.cir.22.4.598
11. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and Sex Estimation Using Artificial Intelligence from Standard 12-Lead ECGs. *Circ Arrhythmia Electrophysiol*. 2019;12(9):1-11. doi:10.1161/CIRCEP.119.007284
12. Zhang Y, Post WS, Blasco-Colmenares E, Dalal D, Tomaselli GF, Guallara E. Electrocardiographic QT interval and mortality: A meta-analysis. *Epidemiology*. 2011;22(5):660-670. doi:10.1097/EDE.0b013e318225768b
13. Noseworthy PA, Peloso GM, Hwang SJ, et al. QT interval and long-term mortality risk in the Framingham heart study. *Ann Noninvasive Electrocardiol*. 2012;17(4):340-348. doi:10.1111/j.1542-474X.2012.00535.x
14. Badheka AO, Singh V, Patel NJ, et al. QRS duration on electrocardiography and cardiovascular mortality (from the national health and nutrition examination survey - III). *Am J Cardiol*. 2013;112(5):671-677. doi:10.1016/j.amjcard.2013.04.040
15. Porthan K, Viitasalo M, Jula A, et al. Predictive value of electrocardiographic QT interval and T-wave morphology parameters for all-cause and cardiovascular mortality in a general population sample. *Hear Rhythm*. 2009;6(8):1202-1208.e1. doi:10.1016/j.hrthm.2009.05.006
16. Kalidas V, Tamil L. Real-time QRS detector using stationary wavelet transform for automated ECG analysis. Proc - *2017*

IEEE 17th Int Conf Bioinforma Bioeng BIBE 2017. 2017;2018-Janua(October):457-461. doi:10.1109/BIBE.2017.00-12

17. Cook S, Malyutina S, Kudryavtsev A V., et al. Know your heart: Rationale, design and conduct of a cross-sectional study of cardiovascular structure, function and risk factors in 4500 men and women aged 35-69 years from two russian cities, 2015-18 [version 3; referees: 3 approved]. *Wellcome Open Res*. 2018;3:1-29. doi:10.12688/wellcomeopenres.14619.3
18. Scheltens T, De Beus MF, Hoes AW, et al. The potential yield of ECG screening of hypertensive patients: The Utrecht Health Project. *J Hypertens*. 2010;28(7):1527-1533. doi:10.1097/HJH.0b013e328339f95c
19. Grobbee DE, Hoes AW, Verheij TJM, Schrijvers AJP, Van Ameijden EJC, Numans ME. The Utrecht Health Project: Optimization of routine healthcare data for research. *Eur J Epidemiol*. 2005;20(3):285-290. doi:10.1007/s10654-004-5689-2
20. Oord A van den, Dieleman S, Zen H, et al. WaveNet: A Generative Model for Raw Audio. Published online 2016:1-15. http://arxiv.org/abs/1609.03499
21. Van De Leur RR, Taha K, Bos MN, et al. Discovering and Visualizing Disease-Specific Electrocardiogram Features Using Deep Learning: Proof-of-Concept in Phospholamban Gene Mutation Carriers. *Circ Arrhythmia Electrophysiol*. 2021;(February):138-147. doi:10.1161/CIRCEP.120.009056
22. Bos MN, Van De Leur RR, Vranken JF, et al. Automated Comprehensive Interpretation of 12-lead Electrocardiograms Using Pre-trained Exponentially Dilated Causal Convolutional Neural Networks. *Comput Cardiol* (2010). 2020;2020-Septe:10-13. doi:10.22489/CinC.2020.253
23. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *IEEE International Conference on Computer Vision*. ; 2017:2999-3007. doi:10.1109/ICCV.2017.324
24. Kingma DP, Ba JL. Adam: A method for stochastic optimization. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc*. Published online 2015:1-15.
25. Steiner B, Devito Z, Chintala S, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems*. ; 2019.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis*. 2020;128(2):336-359. doi:10.1007/s11263-019-01228-7
27. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. ; 2018:839-847. doi:10.1109/WACV.2018.00097
28. Imai K, Keele L, Tingley D. A General Approach to Causal Mediation Analysis. *Psychol Methods*. 2010;15(4):309-334. doi:10.1037/a0020761
29. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation*. 2015;131(2):211-219. doi:10.1161/CIRCULATIONAHA.114.014508
30. Coronel R, Potse M, Haïssaguerre M, et al. Why Ablation of Sites With Purkinje Activation Is Antiarrhythmic: The Interplay Between Fast Activation and Arrhythmogenesis. *Front Physiol*. 2021;12(March). doi:10.3389/fphys.2021.648396
31. Reunanen A, Karjalainen J, Ristola P, Heliovaara M, Knekt P, Aromaa A. Heart rate and mortality. *J Intern Med*. 2000;247(2):231-239. doi:10.1046/j.1365-2796.2000.00602.x
32. Alhalabi L, Singleton MJ, Oseni AO, Shah AJ, Zhang ZM, Soliman EZ. Relation of

Higher Resting Heart Rate to Risk of Cardiovascular Versus Noncardiovascular Death. *Am J Cardiol*. 2017;119(7):1003-1007. doi:10.1016/j.amjcard.2016.11.059
33. Raisi-Estabragh Z, Cooper J, Judge R, et al. Age, sex and disease-specific associations between resting heart rate and cardiovascular mortality in the UK BIOBANK. *PLoS One*. 2020;15(5):1-14. doi:10.1371/journal.pone.0233898
34. Kannel WB, Kannel C, Paffenbarger RS, Cupples LA. Heart rate and cardiovascular mortality: The Framingham study. *Am Heart J*. 1987;113(6):1489-1494. doi:10.1016/0002-8703(87)90666-1
35. Seccareccia F, Pannozzo F, Dima F, et al. Heart rate as a predictor of mortality: The MATISS project. *Am J Public Health*. 2001;91(8):1258-1263. doi:10.2105/AJPH.91.8.1258
36. Vitale C, Fini M, Spoletini I, Lainscak M, Seferovic P, Rosano GM. Under-representation of elderly and women in clinical trials. *Int J Cardiol*. 2017;232:216-221. doi:10.1016/j.ijcard.2017.01.018
37. Pilote L, Raparelli V. Participation of Women in Clinical Trials: Not Yet Time to Rest on Our Laurels. *J Am Coll Cardiol*. 2018;71(18):1970-1972. doi:10.1016/j.jacc.2018.02.069

SUPPLEMENTARY MATERIALS

Supplementary Methods

ACQUISITION OF ECG AND CALCULATION OF ECG VARIABLES IN UMCU DATASET

Conventional ECG parameters, such as ventricular rate, PR interval, QRS duration, QT interval and the frontal R-wave axis were extracted from the MUSE ECG system (MUSE version 8, GE Healthcare, Chicago, IL, USA). QRS- and T-wave minimum, maximum and peak-to-peak voltages were extracted from the median beats per lead using in-house software. The Sokolow-Lyon and Cornell voltage were calculated by adding, respectively, the S-wave voltage in V1 to the maximum of the R-wave voltages in V5 and V6[1] and the S-wave voltage in V3 to the R-wave voltage in aVL.[2] This resulted in a total of 83 conventional ECG variables.

ACQUISITION OF ECG AND CALCULATION OF ECG VARIABLES IN THE KNOW-YOUR-HEART DATASET

All ECGs in the Know-Your-Heart (KYH) dataset were recorded using Cardiax devices (IMED, Hungary) at 500Hz and had differing durations. ECGs shorter than 10 seconds were excluded and ECGs longer than 10 seconds were truncated to the first 10 seconds. Interpretation of the ECG and extraction of conventional ECG parameters was performed with the University of Glasgow ECG analysis program.[3] The ECGs in the validation set were transformed into median beats using the same software as used in the UMCU dataset. Only the baseline normal or borderline normal ECGs, as assessed by the University of Glasgow ECG analysis program, were included in this study.

INCORPORATION OF GUIDED GRAD-CAM IN DNN

The final convolutional layer in the DNN was used for the visualization and the absolute values of the Guided Grad-CAM outputs were obtained from this layer. The visualization was constructed as follows: (1) the median ECG beats and normalized Guided Grad-CAM maps were aligned temporally by the onset of the P-wave, R-wave peak and offset of the T-wave, (2) the mean and standard deviation of the ECG signal were derived per sex in the temporal dimension, (3) the proportion of the Guided Grad-CAM maps above a threshold were derived per timepoint and plotted as a heatmap. The threshold for an important feature was determined by visual inspection.

References

1. Sokolow M, Lyon TP. Electrocardiographic patterns of ventricular hypertrophy as obtained by unipolar precordial and limb leads. *Am J Med*. 1947;2(6):656. doi:10.1016/0002-9343(47)90055-7
2. Casale PN, Devereux RB, Alonso DR, Campo E, Kligfield P. Improved sex-specific criteria of left ventricular hypertrophy for clinical and computer interpretation of electrocardiograms: Validation with autopsy findings. *Circulation*. 1987;75(3):565-572. doi:10.1161/01.CIR.75.3.565
3. Macfarlane PW, Devine B, Clark E. The University of Glasgow (Uni-G) ECG analysis program. *Comput Cardiol*. 2005;32:451-454. doi:10.1109/CIC.2005.1588134

Supplementary table 1 Distribution of ECG variables between correctly classified ECGs in the UMCU internal validation dataset and their misclassified biological counterparts. All values are displayed as median [IQR].

	Overall	Correctly classified	Misclassified	Overall	Correctly classified	Misclassified
n	31,328	26,344	4984	31,260	30,008	1252
VentricularRate	68 [60-79]	67 [59-77]	74 [64-86]	71 [63-81]	71 [63-81]	69 [61-80]
PRInterval	158 [144-174]	158 [144-174]	154 [138-170]	150 [136-166]	150 [136-166]	158 [144-174]
QRSDuration	96 [88-102]	96 [90-102]	90 [84-96]	86 [80-94]	86 [80-92]	92 [86-100]
QTInterval	388 [366-410]	388 [366-410]	382 [356-408]	386 [364-408]	386 [364-408]	388 [366-412]
QTCorrected	413 [401-429]	411 [400-427]	422 [407-439]	419 [406-436]	419 [406-436]	416.50 [405-435]
S-wave amplitude I	-0.09 [-0.16- -0.04]	-0.09 [-0.16- -0.05]	-0.07 [-0.14- -0.03]	-0.07 [-0.12- -0.03]	-0.07 [-0.12- -0.03]	-0.07 [-0.13- -0.03]
R-wave amplitude I	0.71 [0.52-0.92]	0.71 [0.53-0.92]	0.69 [0.50-0.92]	0.69 [0.50-0.92]	0.70 [0.50-0.92]	0.66 [0.48-0.86]
QRS Peak-to-peak I	0.82 [0.63-1.04]	0.82 [0.64-1.04]	0.80 [0.60-1.02]	0.78 [0.59-1.01]	0.78 [0.59-1.01]	0.75 [0.56-0.96]
Minimum amplitude ST-segment I	0 [-0.01-0.01]	0 [-0.01-0.01]	0 [-0.01-0]	0 [-0.01-0]	0 [-0.01-0]	0 [-0.01-0.01]
Maximum amplitude ST-segment I	0.20 [0.14-0.26]	0.20 [0.15-0.27]	0.17 [0.12-0.23]	0.19 [0.14-0.25]	0.19 [0.14-0.25]	0.18 [0.12-0.24]
ST Peak-to-peak I	0.20 [0.14-0.26]	0.20 [0.15-0.26]	0.18 [0.13-0.23]	0.20 [0.15-0.25]	0.20 [0.15-0.25]	0.18 [0.13-0.23]
S-wave amplitude II	-0.10 [-0.20- -0.04]	-0.10 [-0.19- -0.04]	-0.10 [-0.20- -0.04]	-0.10 [-0.19- -0.04]	-0.10 [-0.18- -0.04]	-0.09 [-0.19- -0.04]
R-wave amplitude II	0.79 [0.57-1.05]	0.79 [0.57-1.06]	0.76 [0.56-1]	0.86 [0.65-1.11]	0.86 [0.65-1.11]	0.81 [0.59-1.08]
QRS Peak-to-peak II	0.92 [0.70-1.19]	0.92 [0.70-1.20]	0.89 [0.69-1.14]	0.99 [0.79-1.23]	0.99 [0.79-1.23]	0.93 [0.72-1.20]
Minimum amplitude ST-segment II	0 [-0.01-0.02]	0 [-0.01-0.02]	0 [-0.02-0.01]	0 [-0.02-0.01]	0 [-0.02-0.01]	0 [-0.01-0.02]
Maximum amplitude ST-segment II	0.25 [0.19-0.33]	0.25 [0.19-0.33]	0.24 [0.17-0.31]	0.25 [0.19-0.32]	0.25 [0.19-0.32]	0.25 [0.19-0.34]
ST Peak-to-peak II	0.25 [0.19-0.32]	0.25 [0.19-0.32]	0.24 [0.18-0.31]	0.25 [0.20-0.32]	0.25 [0.20-0.32]	0.25 [0.19-0.32]
S-wave amplitude III	-0.22 [-0.43- -0.10]	-0.22 [-0.43- -0.10]	-0.23 [-0.46- -0.11]	-0.19 [-0.40- -0.09]	-0.19 [-0.40- -0.09]	-0.19 [-0.42- -0.08]
R-wave amplitude III	0.28 [0.13-0.56]	0.28 [0.14-0.57]	0.26 [0.12-0.53]	0.31 [0.14-0.60]	0.31 [0.14-0.60]	0.30 [0.15-0.61]
QRS Peak-to-peak III	0.63 [0.45-0.87]	0.63 [0.45-0.87]	0.62 [0.44-0.86]	0.63 [0.44-0.86]	0.62 [0.44-0.86]	0.64 [0.45-0.87]
Minimum amplitude ST-segment III	-0.02 [-0.05-0]	-0.02 [-0.05-0]	-0.02 [-0.04-0]	-0.02 [-0.04-0]	-0.02 [-0.04-0]	-0.01 [-0.04-0]
Maximum amplitude ST-segment III	0.08 [0.04-0.14]	0.07 [0.04-0.14]	0.08 [0.04-0.14]	0.08 [0.04-0.13]	0.08 [0.04-0.13]	0.09 [0.05-0.16]
ST Peak-to-peak III	0.11 [0.08-0.16]	0.11 [0.08-0.16]	0.11 [0.08-0.16]	0.11 [0.08-0.16]	0.11 [0.08-0.15]	0.12 [0.09-0.18]
S-wave amplitude aVR	-0.75 [-0.91- -0.61]	-0.75 [-0.91- -0.61]	-0.72 [-0.87- -0.58]	-0.77 [-0.93- -0.63]	-0.78 [-0.93- -0.63]	-0.74 [-0.91- -0.59]

	Overall	Correctly classified	Misclassified	Overall	Correctly classified	Misclassified
QRSMax aVR	0.08 [0.04-0.14]	0.08 [0.04-0.14]	0.07 [0.03-0.13]	0.06 [0.03-0.11]	0.06 [0.03-0.11]	0.06 [0.03-0.12]
QRS Peak-to-peak aVR	0.84 [0.70-1]	0.85 [0.70-1.01]	0.81 [0.67-0.96]	0.86 [0.71-1.01]	0.86 [0.72-1.01]	0.82 [0.67-0.99]
Minimum amplitude ST-segment aVR	-0.23 [-0.29- -0.17]	-0.23 [-0.29- -0.18]	-0.21 [-0.26- -0.15]	-0.22 [-0.28- -0.17]	-0.22 [-0.28- -0.17]	-0.22 [-0.28- -0.16]
Maximum amplitude ST-segment aVR	0 [-0.02-0.01]	-0.01 [-0.02-0.01]	0 [-0.01-0.01]	0 [-0.01-0.01]	0 [-0.01-0.01]	0 [-0.01-0.01]
ST Peak-to-peak aVR	0.22 [0.17-0.28]	0.22 [0.18-0.28]	0.21 [0.16-0.26]	0.22 [0.18-0.28]	0.22 [0.18-0.28]	0.21 [0.17-0.27]
S-wave amplitude aVL	-0.12 [-0.24- -0.07]	-0.13 [-0.24- -0.07]	-0.11 [-0.21- -0.06]	-0.10 [-0.21- -0.05]	-0.10 [-0.21- -0.05]	-0.11 [-0.23- -0.06]
R-wave amplitude aVL	0.39 [0.20-0.62]	0.40 [0.20-0.62]	0.39 [0.19-0.63]	0.35 [0.16-0.58]	0.35 [0.16-0.58]	0.34 [0.15-0.57]
QRS Peak-to-peak aVL	0.57 [0.40-0.78]	0.58 [0.41-0.79]	0.55 [0.38-0.77]	0.51 [0.35-0.72]	0.51 [0.35-0.72]	0.51 [0.36-0.72]
Minimum amplitude ST-segment aVL	-0.01 [-0.03-0]	-0.01 [-0.03-0]	-0.02 [-0.04- -0.01]	-0.02 [-0.03- -0.01]	-0.02 [-0.03- -0.01]	-0.02 [-0.04- -0.01]
Maximum amplitude ST-segment aVL	0.08 [0.03-0.14]	0.08 [0.04-0.14]	0.06 [0.03-0.11]	0.08 [0.03-0.12]	0.08 [0.04-0.12]	0.06 [0.02-0.11]
ST Peak-to-peak aVL	0.10 [0.07-0.15]	0.10 [0.07-0.15]	0.09 [0.06-0.13]	0.10 [0.07-0.14]	0.10 [0.07-0.14]	0.09 [0.06-0.12]
S-wave amplitude aVF	-0.11 [-0.22- -0.05]	-0.11 [-0.21- -0.05]	-0.12 [-0.24- -0.06]	-0.11 [-0.21- -0.05]	-0.11 [-0.21- -0.05]	-0.10 [-0.22- -0.05]
R-wave amplitude aVF	0.47 [0.25-0.76]	0.47 [0.26-0.76]	0.46 [0.24-0.72]	0.54 [0.32-0.82]	0.55 [0.32-0.82]	0.51 [0.29-0.79]
QRS Peak-to-peak aVF	0.62 [0.43-0.89]	0.62 [0.43-0.89]	0.63 [0.43-0.87]	0.70 [0.50-0.95]	0.70 [0.50-0.95]	0.66 [0.47-0.95]
Minimum amplitude ST-segment aVF	0 [-0.02-0.01]	0 [-0.02-0.01]	0 [-0.02-0.01]	-0.01 [-0.02-0.01]	-0.01 [-0.02-0.01]	0 [-0.02-0.01]
Maximum amplitude ST-segment aVF	0.16 [0.10-0.22]	0.16 [0.10-0.22]	0.15 [0.10-0.21]	0.16 [0.11-0.22]	0.16 [0.11-0.22]	0.17 [0.11-0.24]
ST Peak-to-peak aVF	0.16 [0.11-0.22]	0.16 [0.11-0.22]	0.16 [0.11-0.22]	0.16 [0.12-0.22]	0.16 [0.12-0.22]	0.17 [0.12-0.24]
S-wave amplitude V1	-0.84 [-1.10- -0.61]	-0.85 [-1.11- -0.62]	-0.80 [-1.05- -0.59]	-0.86 [-1.10- -0.65]	-0.86 [-1.10- -0.65]	-0.88 [-1.16- -0.65]
R-wave amplitude V1	0.18 [0.10-0.28]	0.19 [0.11-0.29]	0.16 [0.09-0.26]	0.16 [0.09-0.25]	0.16 [0.09-0.25]	0.17 [0.10-0.26]
QRS Peak-to-peak V1	1.03 [0.79-1.34]	1.05 [0.80-1.35]	0.98 [0.73-1.26]	1.04 [0.81-1.31]	1.04 [0.81-1.31]	1.07 [0.81-1.38]
Minimum amplitude ST-segment V1	-0.01 [-0.04-0]	-0.01 [-0.04-0.01]	-0.02 [-0.06-0]	-0.04 [-0.09- -0.01]	-0.04 [-0.09- -0.01]	-0.02 [-0.06-0]
Maximum amplitude ST-segment V1	0.10 [0.05-0.19]	0.11 [0.05-0.20]	0.07 [0.03-0.13]	0.05 [0.03-0.10]	0.05 [0.03-0.10]	0.08 [0.04-0.17]
ST Peak-to-peak V1	0.14 [0.09-0.20]	0.14 [0.09-0.21]	0.12 [0.08-0.17]	0.12 [0.08-0.16]	0.12 [0.08-0.16]	0.13 [0.09-0.19]
S-wave amplitude V2	-1.19 [-1.57- -0.86]	-1.21 [-1.60- -0.88]	-1.08 [-1.42- -0.78]	-1.04 [-1.36- -0.78]	-1.04 [-1.35- -0.78]	-1.15 [-1.52- -0.82]
R-wave amplitude V2	0.53 [0.33-0.78]	0.54 [0.34-0.79]	0.46 [0.28-0.70]	0.41 [0.26-0.61]	0.41 [0.26-0.61]	0.45 [0.26-0.72]
QRS Peak-to-peak V2	1.76 [1.36-2.24]	1.79 [1.39-2.27]	1.59 [1.21-2.03]	1.49 [1.17-1.88]	1.49 [1.17-1.87]	1.65 [1.29-2.09]
Minimum amplitude ST-segment V2	0.02 [0-0.04]	0.02 [0-0.04]	0 [-0.02-0.03]	0 [-0.02-0.02]	0 [-0.02-0.02]	0.01 [-0.01-0.03]
Maximum amplitude ST-segment V2	0.51 [0.35-0.68]	0.53 [0.38-0.71]	0.36 [0.24-0.51]	0.32 [0.20-0.45]	0.32 [0.20-0.45]	0.44 [0.29-0.60]

ST Peak-to-peak V2	0.49 [0.34-0.65]	0.51 [0.37-0.68]	0.36 [0.24-0.50]	0.32 [0.21-0.45]	0.32 [0.21-0.44]	0.43 [0.29-0.58]
S-wave amplitude V3	-0.95 [-1.28- -0.66]	-0.97 [-1.30- -0.68]	-0.86 [-1.18- -0.60]	-0.75 [-1.02- -0.51]	-0.75 [-1.02- -0.51]	-0.93 [-1.27- -0.63]
R-wave amplitude V3	0.92 [0.62-1.29]	0.95 [0.65-1.32]	0.79 [0.52-1.12]	0.68 [0.44-1]	0.68 [0.44-0.99]	0.83 [0.52-1.26]
QRS Peak-to-peak V3	1.93 [1.55, 2.37]	1.98 [1.60, 2.42]	1.71 [1.34, 2.12]	1.49 [1.17, 1.88]	1.48 [1.16, 1.86]	1.85 [1.45, 2.29]
Minimum amplitude ST-segment V3	0.02 [0-0.05]	0.02 [0-0.05]	0 [-0.02-0.02]	0 [-0.02-0.02]	0 [-0.02-0.02]	0.01 [-0.02-0.03]
Maximum amplitude ST-segment V3	0.55 [0.40-0.73]	0.58 [0.42-0.76]	0.41 [0.27-0.55]	0.37 [0.25-0.51]	0.37 [0.25-0.51]	0.49 [0.35-0.67]
ST Peak-to-peak V3	0.53 [0.39-0.70]	0.56 [0.41-0.72]	0.40 [0.29-0.54]	0.38 [0.26-0.51]	0.37 [0.26-0.50]	0.49 [0.35-0.65]
S-wave amplitude V4	-0.56 [-0.81- -0.35]	-0.57 [-0.83- -0.35]	-0.51 [-0.75- -0.31]	-0.38 [-0.59- -0.21]	-0.38 [-0.58- -0.21]	-0.52 [-0.79- -0.30]
R-wave amplitude V4	1.45 [1.08-1.86]	1.51 [1.14-1.92]	1.16 [0.83-1.52]	1.03 [0.74-1.39]	1.02 [0.73-1.37]	1.38 [1-1.82]
QRS Peak-to-peak V4	2.05 [1.66-2.50]	2.12 [1.74-2.56]	1.71 [1.36-2.09]	1.46 [1.15-1.82]	1.44 [1.14-1.80]	1.96 [1.54-2.41]
Minimum amplitude ST-segment V4	0 [-0.02-0.03]	0.01 [-0.01-0.03]	-0.01 [-0.03-0.01]	-0.01 [-0.03-0.01]	-0.01 [-0.03-0]	0 [-0.02-0.02]
Maximum amplitude ST-segment V4	0.45 [0.31-0.61]	0.47 [0.34-0.63]	0.33 [0.22-0.46]	0.29 [0.20-0.40]	0.28 [0.20-0.40]	0.41 [0.29-0.57]
ST Peak-to-peak V4	0.44 [0.32-0.59]	0.46 [0.34-0.61]	0.34 [0.24-0.46]	0.30 [0.21-0.41]	0.29 [0.21-0.40]	0.41 [0.30-0.56]
S-wave amplitude V5	-0.24 [-0.40- -0.13]	-0.24 [-0.40- -0.13]	-0.24 [-0.41- -0.12]	-0.18 [-0.31- -0.09]	-0.18 [-0.30- -0.09]	-0.20 [-0.35- -0.10]
R-wave amplitude V5	1.47 [1.16-1.83]	1.52 [1.21-1.87]	1.23 [0.94-1.57]	1.18 [0.91-1.48]	1.17 [0.91-1.47]	1.40 [1.10-1.74]
QRS Peak-to-peak V5	1.76 [1.42-2.13]	1.80 [1.47-2.17]	1.51 [1.21-1.86]	1.39 [1.12-1.70]	1.38 [1.12-1.69]	1.64 [1.34-2]
Minimum amplitude ST-segment V5	0 [-0.02-0.02]	0 [-0.02-0.02]	-0.01 [-0.02-0.01]	-0.01 [-0.02-0]	-0.01 [-0.02-0]	0 [-0.02-0.01]
Maximum amplitude ST-segment V5	0.34 [0.24-0.47]	0.36 [0.25-0.48]	0.28 [0.19-0.39]	0.27 [0.19-0.37]	0.27 [0.19-0.36]	0.32 [0.23-0.43]
ST Peak-to-peak V5	0.35 [0.25-0.46]	0.36 [0.26-0.47]	0.29 [0.20-0.39]	0.28 [0.20-0.37]	0.28 [0.20-0.37]	0.33 [0.23-0.42]
S-wave amplitude V6	-0.10 [-0.17- -0.05]	-0.10 [-0.17- -0.05]	-0.10 [-0.17- -0.05]	-0.08 [-0.14- -0.04]	-0.08 [-0.14- -0.04]	-0.09 [-0.14- -0.04]
R-wave amplitude V6	1.11 [0.87-1.39]	1.14 [0.89-1.41]	1 [0.77-1.27]	1.03 [0.81-1.27]	1.03 [0.81-1.27]	1.06 [0.82-1.33]
QRS Peak-to-peak V6	1.23 [0.98-1.53]	1.25 [1-1.55]	1.12 [0.88-1.41]	1.13 [0.90-1.38]	1.13 [0.90-1.38]	1.16 [0.92-1.45]
Minimum amplitude ST-segment V6	0 [-0.01-0.02]	0 [-0.01-0.02]	0 [-0.02-0.01]	0 [-0.02-0]	0 [-0.02-0]	0 [-0.01-0.01]
Maximum amplitude ST-segment V6	0.25 [0.17-0.34]	0.25 [0.18-0.35]	0.22 [0.15-0.30]	0.23 [0.17-0.30]	0.23 [0.17-0.30]	0.23 [0.17-0.31]
ST Peak-to-peak V6	0.25 [0.18-0.33]	0.25 [0.18-0.34]	0.22 [0.16-0.30]	0.23 [0.18-0.30]	0.23 [0.18-0.30]	0.23 [0.18-0.31]
Sokolow-Lyon voltage	1.99 [1.60-2.41]	2.02 [1.63-2.44]	1.83 [1.47-2.24]	1.91 [1.55-2.30]	1.91 [1.55-2.30]	1.98 [1.59-2.41]
Cornell voltage	1.40 [1.06-1.75]	1.41 [1.08-1.77]	1.31 [0.98-1.66]	1.16 [0.86-1.49]	1.15 [0.85-1.48]	1.32 [0.99-1.70]
Sokolow-Lyon product	188.1 [148.7-233.0]	192.8 [153.2-237.5]	163.4 [130.5-202.3]	165.9 [132.2-202.9]	165.2 [131.8-202.0]	182.2 [145.9-227.3]
Cornell product	132.2 [97.8-170.8]	135.3 [100.5-174.0]	117.0 [85.5-152.9]	100.2 [72.3-132.2]	99.4 [71.8-131.1]	121.9 [88.9-159.3]

Supplementary table 2 Proportion mediated with 95%- confidence intervals per ECG variable in the relation of sex classification by the DNN and survival, analysis performed stratified by sex.

Mediators	Average proportion mediated by ECG variable in men (95% CI)	Average proportion mediated by ECG variable in women (95% CI)	Mediators	Average proportion mediated by ECG variable in men (95% CI)	Average proportion mediated by ECG variable in women (95% CI)
PRInterval	0.09 (0.06, 0.14)	-0.07 (-0.14,-0.04)	QRS Peak-to-peak I	0.03 (0.02, 0.05)	0.06 (0.03, 0.13)
QRSDuration	0.18 (0.11, 0.26)	-0.25 (-0.52,-0.15)	Maximum amplitude ST-segment I	0.16 (0.11, 0.22)	0.09 (0.05, 0.18)
QTCorrected	0.21 (0.16, 0.29)	-0.05 (-0.11,-0.02)	Minimum amplitude ST-segment I	0.10 (0.06, 0.14)	-0.04 (-0.08,-0.02)
QTInterval	0.17 (0.12, 0.23)	-0.04 (-0.09,-0.01)	ST Peak-to-peak I	0.12 (0.08, 0.16)	0.13 (0.08, 0.25)
SLProduct	0.08 (0.04, 0.13)	-0.06 (-0.13,-0.01)	R-wave amplitude V1	0.01 (0.00, 0.03)	-0.03 (-0.07,-0.01)
SLVoltage	0.02 (0.00, 0.05)	-0.01 (-0.03, 0.01)	S-wave amplitude V1	-0.01 (-0.02, 0.00)	0.02 (0.01, 0.06)
VentricularRate	0.39 (0.31, 0.54)	-0.08 (-0.18,-0.03)	QRS Peak-to-peak V1	0.00 (-0.01, 0.01)	0.02 (0.00, 0.04)
R-wave amplitude aVF	0.00 (0.00, 0.00)	0.00 (-0.01, 0.01)	Maximum amplitude ST-segment V1	-0.05 (-0.10,-0.02)	0.18 (0.10, 0.35)
S-wave amplitude aVF	-0.01 (-0.02, 0.00)	0.00 (0.00, 0.01)	Minimum amplitude ST-segment V1	0.02 (-0.02, 0.05)	0.03 (0.01, 0.08)
QRS Peak-to-peak aVF	0.00 (0.00, 0.01)	0.00 (-0.01, 0.00)	ST Peak-to-peak V1	-0.05 (-0.08,-0.02)	0.12 (0.06, 0.24)
Maximum amplitude ST-segment aVF	0.00 (-0.01, 0.00)	0.03 (0.01, 0.07)	R-wave amplitude V2	0.06 (0.04, 0.09)	-0.03 (-0.08,-0.01)
Minimum amplitude ST-segment aVF	-0.01 (-0.02, 0.00)	0.04 (0.02, 0.09)	S-wave amplitude V2	-0.01 (-0.03, 0.01)	0.08 (0.04, 0.16)
ST Peak-to-peak aVF	0.00 (0.00, 0.01)	0.01 (0.00, 0.04)	QRS Peak-to-peak V2	0.04 (0.01, 0.07)	0.04 (0.00, 0.11)
R-wave amplitude aVL	0.01 (0.00, 0.02)	0.04 (0.02, 0.08)	Maximum amplitude ST-segment V2	0.18 (0.11, 0.28)	0.00 (-0.08, 0.07)
S-wave amplitude aVL	-0.01 (-0.02, 0.00)	0.02 (-0.01, 0.05)	Minimum amplitude ST-segment V2	0.13 (0.08, 0.20)	-0.01 (-0.07, 0.02)
QRS Peak-to-peak aVL	0.02 (0.01, 0.03)	0.01 (0.00, 0.03)	ST Peak-to-peak V2	0.15 (0.09, 0.24)	0.00 (-0.07, 0.08)
Maximum amplitude ST-segment aVL	0.12 (0.08, 0.16)	0.09 (0.05, 0.19)	R-wave amplitude V3	0.10 (0.06, 0.15)	-0.02 (-0.07, 0.02)
Minimum amplitude ST-segment aVL	0.08 (0.06, 0.12)	0.07 (0.04, 0.13)	S-wave amplitude V3	-0.03 (-0.06,-0.01)	0.13 (0.07, 0.24)
ST Peak-to-peak aVL	0.06 (0.04, 0.08)	0.04 (0.02, 0.09)	QRS Peak-to-peak V3	0.05 (0.01, 0.10)	0.09 (0.03, 0.21)
R-wave amplitude aVR	0.00 (0.00, 0.00)	-0.01 (-0.03, 0.00)	Maximum amplitude ST-segment V3	0.10 (0.03, 0.17)	0.04 (-0.04, 0.12)
S-wave amplitude aVR	0.02 (0.01, 0.04)	0.07 (0.03, 0.13)	Minimum amplitude ST-segment V3	0.12 (0.07, 0.19)	0.04 (0.00, 0.11)
QRS Peak-to-peak aVR	0.03 (0.01, 0.04)	0.06 (0.03, 0.12)	ST Peak-to-peak V3	0.07 (0.00, 0.14)	0.02 (-0.05, 0.11)
Maximum amplitude ST-segment aVR	0.05 (0.02, 0.09)	0.01 (-0.02, 0.04)	R-wave amplitude V4	0.16 (0.10, 0.24)	-0.09 (-0.21,-0.02)

Minimum amplitude ST-segment aVR	0.09 (0.06, 0.13)	0.01 (-0.02, 0.04)	S-wave amplitude V4	-0.03 (-0.06,-0.01)	0.13 (0.07, 0.26)
ST Peak-to-peak aVR	0.06 (0.04, 0.09)	0.04 (0.01, 0.09)	QRS Peak-to-peak V4	0.10 (0.03, 0.17)	0.03 (-0.07, 0.15)
CornellProduct	0.03 (0.00, 0.06)	-0.01 (-0.07, 0.04)	Maximum amplitude ST-segment V4	0.02 (-0.05, 0.09)	0.05 (-0.03, 0.17)
CornellVoltage	0.00 (-0.02, 0.02)	0.02 (-0.01, 0.06)	Minimum amplitude ST-segment V4	0.06 (0.02, 0.11)	0.07 (0.02, 0.14)
R-wave amplitude III	0.00 (-0.01, 0.00)	0.01 (0.00, 0.03)	ST Peak-to-peak V4	0.00 (-0.06, 0.06)	0.03 (-0.06, 0.12)
S-wave amplitude III	0.00 (-0.01, 0.00)	0.01 (0.00, 0.03)	R-wave amplitude V5	0.15 (0.09, 0.23)	-0.09 (-0.20,-0.03)
QRS Peak-to-peak III	0.00 (0.00, 0.00)	0.00 (-0.01, 0.00)	S-wave amplitude V5	0.00 (-0.01, 0.00)	0.03 (0.01, 0.06)
Maximum amplitude ST-segment III	0.03 (0.02, 0.04)	0.09 (0.05, 0.19)	QRS Peak-to-peak V5	0.10 (0.05, 0.17)	-0.05 (-0.14, 0.01)
Minimum amplitude ST-segment III	0.03 (0.02, 0.05)	0.02 (0.00, 0.06)	Maximum amplitude ST-segment V5	0.06 (0.02, 0.12)	-0.03 (-0.10, 0.02)
ST Peak-to-peak III	0.00 (0.00, 0.01)	0.06 (0.03, 0.11)	Minimum amplitude ST-segment V5	0.05 (0.02, 0.09)	0.03 (0.01, 0.07)
R-wave amplitude II	0.00 (0.00, 0.01)	0.01 (0.00, 0.02)	ST Peak-to-peak V5	0.05 (0.01, 0.09)	-0.04 (-0.10, 0.01)
S-wave amplitude II	0.00 (-0.01, 0.00)	0.00 (0.00, 0.01)	R-wave amplitude V6	0.06 (0.03, 0.10)	-0.03 (-0.08,-0.01)
QRS Peak-to-peak II	0.00 (-0.01, 0.00)	0.01 (0.00, 0.03)	S-wave amplitude V6	0.01 (0.00, 0.02)	0.00 (0.00, 0.01)
Maximum amplitude ST-segment II	0.01 (0.00, 0.02)	-0.01 (-0.03, 0.00)	QRS Peak-to-peak V6	0.04 (0.01, 0.07)	-0.03 (-0.07,-0.01)
Minimum amplitude ST-segment II	0.00 (-0.02, 0.03)	0.04 (0.01, 0.09)	Maximum amplitude ST-segment V6	0.05 (0.02, 0.08)	-0.03 (-0.07,-0.01)
ST Peak-to-peak II	0.01 (0.00, 0.02)	-0.01 (-0.02, 0.00)	Minimum amplitude ST-segment V6	0.05 (0.02, 0.09)	0.02 (-0.01, 0.05)
R-wave amplitude I	0.03 (0.01, 0.04)	0.08 (0.04, 0.15)	ST Peak-to-peak V6	0.03 (0.01, 0.06)	-0.02 (-0.06,-0.01)
S-wave amplitude I	0.00 (-0.01, 0.01)	-0.02 (-0.05, 0.00)			

Chapter 10

General discussion and future perspectives

There is a worldwide growing demand for cardiovascular healthcare. This is due to an increasing prevalence of cardiovascular disease (CVD)[1,2], but also due to improved diagnostics and therapeutics[3]. This has converted CVD from a lethal into a chronic condition, which makes the amount of regular care data on CVD abundantly present in contemporary healthcare. The concurrent shift from paper reporting to electronic health records (EHR) has increased the availability and quality of medical data and data management.[4] These characteristics make EHR more accessible and better suitable to perform medical research on. On top, the chronicity of CVD paved the way for self-monitoring by patients. Self-monitoring is for instance the use of telemonitoring of blood pressure and ECGs at the comfort of the patient's own home, instead of paying regular visits to a physician. It positions the patient in control of their own care and gives the treating cardiologist a consulting role, helping out when required. This way of providing care has shown to generate accurate and reliable data and improving medical conditions.[5,6]

The increasing prevalence of CVD, CVD turning into a chronic condition, and the digitalization of health records and self-monitoring generate an increased amount of regular care data that is instantly available in a digital way. This broadens the horizons of cardiovascular research and gives the opportunity to complement data and results from clinical studies. Clinical studies, and specifically randomized controlled trials, face the burden of strict in- and exclusion criteria, which reduces generalizability and hampers a universal application of clinical guidelines. Real world data, including regular care data has the potential to increase knowledge of presentation of CVD in populations that have long been underrepresented in cardiovascular clinical trials, i.e. women[7], ethnic minorities and patients with multimorbidity[8,9]. It also creates the opportunity to study the clinical implementation and performance of clinical guidelines in the actual population that these guidelines are intended for.

In addition, these new types of data ask for new ways of analysis and interpretation, as the cardiologist and care personnel will be faced with a never-ending amount of cardiovascular data to be examined. In addition, data may hold information that cannot be seen by eye. Artificial Intelligence (AI) has the potential to automatically analyse large amounts of unstructured data and generate conclusions from these data. This is in contrast to the traditional statistical methods, that require structured data to perform analyses upon. Nonetheless, current clinical guidelines heavily rely on data from clinical studies that have been analysed with these classical methods.

The aim of this thesis entitled "Moving from traditional methods towards artificial intelligence in cardiovascular research with regular care data" was to investigate and explore the use of different traditional statistical methods and AI on regular care data. Beyond the exploration of the appropriate use of these methods, this thesis shows how regular care data can be used to provide insight into sex differences in CVD. In this discussion, I

will elaborate in more detail on when application of AI is relevant in the CVD domain. In this reflection, I take into account the different types of data that are present in regular care. As AI has not yet found its way into clinical practice[10–12], I will also zoom in on the hurdles that have to be taken to transition AI from research settings towards clinical cardiology practice.

WHEN TO CHOOSE AI OVER TRADITIONAL STATISTICAL METHODS IN CARDIOVASCULAR RESEARCH WITH REGULAR CARE DATA?

AI has the potential to generate meaningful results from unstructured data, i.e. data that is not systematically registered. However, large amounts of data are required to train AI models to identify patterns and provide meaningful results. Given these characteristics of AI - the ability to turn unstructured data into structured data and the requirement of large datasets - regular care data is a suitable data type for AI methods. Regular care data is abundantly present and consists of both structured, e.g. lab values and vital values, and unstructured data sources, e.g. cardiovascular medical imaging (Chapter 7), cardiovascular electrophysiological signals (Chapter 9) with ECGs being the most frequent, and free clinical text (Chapter 6). Yet, research goals and data quality must be considered before blindly applying AI to a regular care database.

AI is no panacea that turns unstructured data into a clean and comprehensive dataset. Research questions and goals that are suitable for AI should therefore be carefully considered. The current application of AI is 'narrow', which means that it is trained in one particular task that it performs in very well. Examples can be found in studies done using one of the unstructured data sources in cardiovascular regular care. For cardiovascular imaging, a narrow task is, for example, the selection of the appropriate imaging protocol, based on filed information by the requesting physician[13–15] or automatic segmentation and quantification of the left ventricular volume[16]. Narrow tasks performed by AI using clinical notes include the extraction of cardiovascular risk factors from text[17,18] and turning radiology reports into a useful featureset[19]. AI applied for analysis of ECGs varies from diagnosis[20,21] and triage[22] of ECGs to derivation of ECG conduction intervals[23]. However, we show in Chapter 9 that AI can be applied in a broader setting. For instance, it can be used in a research setting to generate new hypotheses. In Chapter 9, we trained a deep neural network to predict sex based on ECG signals, in order to identify sex-specific ECG features that might be associated with mortality. This approach of AI practice showed an association between shorter QRS duration and reduced survival in males and females, which has not been reported before and shows that AI has the possibility to entangle patterns in large amounts of data which cannot be identified by the human brain. AI can thus lead to new hypotheses and discoveries in cardiovascular research.

All the unstructured data types that have been mentioned above – medical imaging, electrophysiological signals and text – cannot be (automatically) analysed with tradition-

al statistical methods, as the amount of data generated is either too large or too complex for these traditional methods. Nevertheless, when unstructured data are turned into structured data, for example in Chapter 8 of this thesis, both AI and traditional statistical methods can be used. In these cases, the researcher has to assess which method is most suitable given the research question, the amount of data samples/inclusions, the number of features, quality of the data or clinical use of the algorithm. Also the purpose of the intended research is important in this matter, as an easily interpretable logistic regression that provides similar outcomes, is likely to be preferred by clinicians and patients over more complex and sophisticated AI methods, as these may be hard to interpret. This was also shown in Chapter 8. In this chapter, we demonstrated that a Lasso logistic regression and a gradient boosting model improved the diagnosis of coronary artery disease in patients, and specifically in women, with chest pain or dyspnoea compared to the existing risk score based on three classical patient characteristics[24]. This study showed that the application of AI on a large structured database improved patient profiling on top of existing risk scores, that are well calibrated.[25] Nonetheless, implementation in clinical practice can be cumbersome, due to the large amount of variables used in these models. Clinicians might still prefer the classical risk score based on three patient characteristics. Not always does AI outperform classical risk scores, i.e. for the prediction of outcomes in atrial fibrillation on top of traditional risk scores[26] and prediction of cardiovascular disease[27]. It is likely that in these cases AI is actually analysing the same high risk features, i.e. age, type of complaints and sex, as are incorporated in risk scores. These scores have been calibrated and already provide a good risk stratification. Each additional feature will only have little incremental value.[28]

In regular care data, data quality is often hampered due to the presence of missing values, which can in some cases be selective and informative. This means that the absence of a variable contains information about the patient.[29] Informative or selective missingness can in some cases be overcome by linkage to other data sources to extract relevant information or to enrich the dataset.[30–33] Informative missingness can also be incorporated in datasets (Chapter 8) with the use of dummy variables.[29] Furthermore, imputation methods are able to deal with missingness[34] (Chapter 3, Chapter 5B and Chapter 8) in a research setting. However, in general practice missing values will remain, which impedes the use of real-time risk calculations in the EHR. Currently, research groups are working on possible solutions using real-time imputation.[35]

Previous examples showed the promise of AI in the current era of healthcare digitalization taking into account research goals and data quality. The ever-growing amounts of data that are generated and digitalized in cardiovascular care by ECG-telemonitoring systems and smartwatches[36], clinical notes of cardiac consults and imaging procedures performed to rule-out disease, just cannot be processed by physicians alone. AI can en-

hance physician's life and workload. Nevertheless, AI is trained to be exceptionally great at one task, e.g. segmentation of the left ventricle on MRI[37] or automatic classification of the ECG[38]. It should be considered as assisting the treating physician. It can therefore never be the substitute of a medical doctor, who is able to interpret the complexity of patient's condition. To show empathy, consider the patient as a whole and have interdisciplinary and interpersonal skills and knowledge is specific to human. Yet, given its great potential in supporting the physician, there is still surprisingly little implementation of AI in clinical practice.

FROM CODE TO… CLINIC: THE DELAY IN AI INNOVATION

The use of AI in healthcare research has significantly increased in recent years[39] which indicates the potential of AI for clinical cardiovascular care. However, at this point clinical implementation of AI models is very limited to almost none. Nevertheless, innovation always takes time and, especially in healthcare due to the responsibility for and effect on a patient's life, innovation is difficult[40,41]. Therefore, the delay in clinical implementation is argumentative. The timely delay between the development of an algorithm in a research environment and the launch of a clinically relevant, efficient and viable AI product has been indicated as the 'AI chasm'.[10] To gain more insight into the 'AI chasm' it is important to evaluate the different hurdles that have to be taken in the trajectory from code to clinic.

Currently, there is a lack of external validation studies for AI models developed in a research setting. This was already shown for radiology, the frontrunning medical field regarding AI, in which only 6% of studies that described development and performance of an AI algorithm for a specific task, performed external validation.[11] To perform external validation of an AI algorithm, an external validation dataset is required.[42] Acquisition of an external validation dataset can be a cumbersome process, due to strict data protection regulations[43] and data architecture requirements to ensure compliance[44]. Development of AI models also requires labelled datasets. This requires intensive manual labelling of samples, i.e. extraction of features from free text[45,46], manual annotation of ECGs or manual segmentation of cardiovascular imaging. Fortunately, the use of open data resources are becoming more popular in the development of AI in healthcare[47], on top of journals that require data availability statements before publication[48]. Yet, when an external dataset is available for validation, this dataset should be aligned to the dataset used for training. For research done with regular care data, and specifically EHR data, this means that the data must be interoperable at national and even international levels. This has proven to be very difficult, although attempts are made to standardize the clinical meaning of variables with i.e. FHIR[49,50], ICD-10 and SNO-MED[51]. Another example of data alignment was presented in Chapter 9. To test model performance on the external datasets, the ECG was converted into a median beat ECG instead of 10-second ECG data.

This was also done to the ECGs in the external datasets, aligning all data and enabling predictions on these data.

The lack of external validation of AI models also decreases chance of widespread implementation in clinical practice due to reduced generalizability. Sex[52] and race are two important characteristics that can lead to non-generalizable AI models. This is specifically applicable to ECG data, as ECG characteristics differ between the sexes and ethnicities[53,54], but also change with increasing age[55,56]. The current lack of reporting the performance of an algorithm in subgroups makes this application of AI prone to bias. Fortunately, to date, no studies proved non-generalizability in AI algorithms in cardiovascular research. Nevertheless, external validation and reporting of the performance of the AI model in different subgroups is necessary to evaluate and proof generalizability and to identify possible limitations in training data or a need for subgroup-specific models. This was done for a model developed to diagnose atrial fibrillation with ECGs[57] and in Chapter 8 of this thesis.

Beyond external validation of AI models, researchers must also have the opportunity to reproduce and replicate AI models from other researchers. Replicable and reproducible research can speed up the implementation process by increasing the validity of AI-based research and findings. Replicability and reproducibility are important aspects of the shift in our research climate towards open science, in which code and data should be openly available to anyone interested. However, research has shown that AI publications in healthcare currently lack the possibility for replication and reproduction[58], in contrast to the widespread uptake of Open Science by the AI community.[47] First, when focusing on reproducibility (i.e. using the original data and code to reach the same results) of AI in healthcare, this is hampered by, on the one hand, the willingness to share models and data due to data protection regulations. On the other hand, some level of randomness is induced during model development and model settings, variables and architecture are not always structurally reported in medical journals.[59] Although not common practice in healthcare, the call for action was picked up as shown by an addendum[60] from Google Health after publication of their AI breast cancer screening tool.[61] This addendum gave more information about data augmentation and optimization of the model parameters. Second, when focusing on replicable (i.e. using a different dataset and code to reach the same results) research, it is important to uncover potentially dangerous biases in AI models.[62] Although no bias has been shown in AI research in the cardiovascular domain to date, examples of sex, gender[52] and racial bias[63] in AI models in healthcare are present. This bias can arise from multiple factors. Bias present in the training dataset might leak into the decisions the model makes. An example thereof is that women experience less classical symptoms during a heart attack[64] and women more often have silent ischemia[65]. If these women are not referred for appropriate care, this is not recorded in the regular

care database. As a consequence, the use of this database for AI applications integrates incorrect referral of women that experience a heart attack. Another source of bias is the use of proxies in AI models to represent an unreported variable. This was shown in a study that predicted an individual's medical expenditures. It showed that, although individuals had the same number of medical expenditures, self-reported black individuals had more chronic illnesses than self-reported white individuals. Therefore, using medical expenditures as a proxy for condition of health leads to an underestimation of the severity of disease in black individuals.[63]

Beyond validation and replicability of models, the value of AI in clinical practice should be evaluated. In a regular care database, this can be done retrospectively by comparison to a traditional statistical risk score or the use of the, for example, net reclassification index[66]. These measures aid in valuing the AI model and the potential added value in the clinic. Although this is not yet common practice, an increasing number of studies are implementing these measures.[67]

Nonetheless, the actual implementation of AI and the effect it has on clinical workflow should be prospectively evaluated. Clinical efficacy is not shown in retrospective trials, but is obtained through implementation of the AI model in a (randomized) controlled trial. Clinical efficacy can be evaluated at six different levels, ranging from technical performance towards societal impact. Van Leeuwen et al. proposed a model for clinical efficacy specifically tailored towards validation of AI models in radiology.[68] This model was adapted from Fryback and Thornbury[69]. However, evidence on clinical efficacy of AI is currently lacking. This was illustrated by an overview of all the AI applications in radiology that already obtained CE-marking – a certification that shows the product has been assessed to meet high safety, health and environment requirements. More than half of these products lacked peer-reviewed evidence on efficacy.[68] This can for example be done by a three-way comparison between performance of the AI algorithm, the clinician's performance and the performance of the AI algorithm in collaboration with or supervised by the clinician.[70] The latter situation has shown supreme performance over each entity individually in some cases.[71] Nonetheless, the design of a clinical trial with AI models depends on the clinical implementation and role of the designed AI model. Supervision of AI models will always be required, due to liability and medicolegal issues. To create a fruitful and optimal cooperation between the AI software and the clinician, physician's trust in the AI is required. To facilitate this trustworthy relation between AI applications and the physicians, also legal issues should be addressed. Currently, decisions that an AI model makes are hard to interpret and explain in detail, creating the idea of a "black box". This type of decision-making makes physicians uncomfortable working with AI and distrusting the technology[72]. It also gives physicians a certain liability to what extent they will follow the outcome of the AI algorithm when its decision differs from the

standard of care.[73] This has recently led to the introduction and expansion of the field of explainable AI[74], which will hopefully increase trust among physicians to start collaborating with AI algorithms.

On top, also the patient is an important stakeholder in the implementation of AI in clinical care. In the end, it is his or her healthcare that is being decided upon by an algorithm. Although this discussion specifically focused on the practical limitations of AI, the role of the patient in this process should be clear. As long as AI is implemented as a black-box in clinical care, shared-decision making is hampered. This reduces the perceived benefit for the patient as no explanation is given with the AI judgments.[75] Also one step before, during development of AI models, patient involvement is required in the possible data collection and use of their (anonymized) healthcare data. An elaborate analysis of the legal and ethical basis of regular care data use and reuse for cardiovascular research is beyond the scope of this thesis.

As a consequence of all different reasons mentioned that delay implementation of AI in clinical practice, it becomes clear that many stakeholders, ranging from data scientists to physicians and UX designers to security officers, must be involved in the development and implementation of AI in regular clinical practice.[76] Development of AI models already takes a significant amount of time, but the behavioural change that must be established in the hospital's environment might even take a longer time.[77] This behavioural change should be a combinative effort of all different stakeholders, and includes different aspects of implementation of AI, e.g. economics, effectivity and efficiency, user interface and design, and patient experience.

To conclude, AI is a very promising and powerful tool to use when the input data is unstructured, illustrated in this thesis by the application of AI on cardiovascular imaging, electrophysiological signals and free text. For more structured databases AI and traditional statistical methods can perform evenly well and the more interpretable and intuitive method should be used. Several data quality requirements must be met before regular care data can be used in either AI or traditional statistical methods, because the end product is only as good as the source data ('garbage in, garbage out'-principle). Code can only be transferred to clinic, when a physician trusts the model, which starts with uncovering of the black box. This also includes the evaluation of potential sex and racial bias in an AI system. In the end, AI will not replace the physician, but it will enhance their daily workflow and allocation of their time, resulting in more time and empathy for the patient to build a sustainable and trustworthy relationship.

REFERENCES

1. Roth GA, Mensah GA, Johnson CO, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol*. 2020;76(25):2982-3021. doi:10.1016/j.jacc.2020.11.010
2. de Boer, A.R.; van Dis, I.; Wimmers, R.H.; Vaartjes, I.; Bots ML. Cijfers- Hart En Vaatziekten in Nederland 2020. *Cent Bur voor Stat en Niv*. Published online 2020.
3. IOM (Institute of Medicine). Reducing the Burden of Cardiovascular Disease: Intervention Approaches. In: Fuster V, Kelly BB, eds. *Promoting Cardiovascular Health in the Developing World: A Challenge to a Acheive Global Health*. The National Academies Press; 2010:185-274.
4. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. *J Med Syst*. 2018;42(11). doi:10.1007/s10916-018-1075-6
5. Paré G, Jaana M, Sicotte C. Systematic Review of Home Telemonitoring for Chronic Diseases: The Evidence Base. *J Am Med Informatics Assoc*. 2007;14(3):269-277. doi:10.1197/jamia.M2270
6. Kolk MZH, Blok S, De Wildt MCC, et al. Patient-reported outcomes in symptom-driven remote arrhythmia monitoring: evaluation of the Dutch HartWacht-telemonitoring programme. *Eur Hear J - Digit Heal*. 2021;2(2):224-230. doi:10.1093/ehjdh/ztab030
7. Pilote L, Raparelli V. Participation of Women in Clinical Trials: Not Yet Time to Rest on Our Laurels. *J Am Coll Cardiol*. 2018;71(18):1970-1972. doi:10.1016/j.jacc.2018.02.069
8. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals. *Jama*. 2007;297(11):1233-1240. doi:10.1001/jama.298.1.39-b
9. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA Intern Med*. 2014;174(11):1868-1870. doi:10.1001/jamainternmed.2014.4758
10. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *npj Digit Med*. 2018;1(1):10-12. doi:10.1038/s41746-018-0048-y
11. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. *Korean J Radiol*. 2019;20(3):405-410. doi:10.3348/kjr.2019.0025
12. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Futur Healthc J*. 2019;6(2):94-98. doi:10.7861/futurehosp.6-2-94
13. Bizzo BC, Almeida RR, Michalski MH, Alkasab TK. Artificial Intelligence and Clinical Decision Support for Radiologists and Referring Providers. *J Am Coll Radiol*. 2019;16(9):1351-1356. doi:10.1016/j.jacr.2019.06.010
14. Kalra A, Chakraborty A, Fine B, Reicher J. Machine Learning for Automation of Radiology Protocols for Quality and Efficiency Improvement. *J Am Coll Radiol*. 2020;17(9):1149-1158. doi:10.1016/j.jacr.2020.03.012
15. Nencka AS, Sherafati M, Goebel T, Tolat P, Koch KM. Deep-learning based Tools for Automated Protocol Definition of Advanced Diagnostic Imaging Exams. Published online 2021:1-15. http://arxiv.org/abs/2106.08963
16. Wang T, Lei Y, Tang H, et al. A learning-based automatic segmentation and quantification method on left ventricle

in gated myocardial perfusion SPECT imaging: A feasibility study. *J Nucl Cardiol*. 2020;27(3):976-987. doi:10.1007/s12350-019-01594-2
17. Groenhof TKJ, Koers LR, Blasse E, et al. Data mining information from electronic health records produced high yield and accuracy for current smoking status. *J Clin Epidemiol*. 2020;118:100-106. doi:10.1016/j.jclinepi.2019.11.006
18. Moon S, Liu S, Scott CG, et al. Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. *Int J Med Inform*. 2019;128(September 2018):32-38. doi:10.1016/j.ijmedinf.2019.05.008
19. Zheng C, Sun BC, Wu YL, et al. Automated abstraction of myocardial perfusion imaging reports using natural language processing. *J Nucl Cardiol*. Published online 2020. doi:10.1007/s12350-020-02401-z
20. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11(1):1-9. doi:10.1038/s41467-020-15432-4
21. Grün D, Rudolph F, Gumpfer N, et al. Identifying Heart Failure in ECG Data With Artificial Intelligence—A Meta-Analysis. *Front Digit Heal*. 2021;2(February):1-7. doi:10.3389/fdgth.2020.584555
22. van de Leur RR, Blom LJ, Gavves E, et al. Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks. *J Am Heart Assoc*. 2020;9(10):e015138. doi:10.1161/JAHA.119.015138
23. Giudicessi JR, Schram M, Bos JM, et al. Artificial Intelligence-Enabled Assessment of the Heart Rate Corrected QT Interval Using a Mobile Electrocardiogram Device. *Circulation*. Published online 2021:1274-1286. doi:10.1161/CIRCULATIONAHA.120.050231
24. Juarez-Orozco LE, Saraste A, Capodanno D, et al. Impact of a decreasing pre-test probability on the performance of diagnostic tests for coronary artery disease. *Eur Heart J Cardiovasc Imaging*. 2019;20(11):1198-1207. doi:10.1093/ehjci/jez054
25. Winther S, Schmidt SE, Rasmussen LD, et al. Validation of the European Society of Cardiology pre-test probability model for obstructive coronary artery disease. *Eur Heart J*. Published online 2020:1-11. doi:10.1093/eurheartj/ehaa755
26. Loring Z, Mehrotra S, Piccini JP, et al. Machine learning does not improve upon traditional regression in predicting outcomes in atrial fibrillation: An analysis of the ORBIT-AF and GARFIELD-AF registries. *Europace*. 2020;22(11):1635-1644. doi:10.1093/europace/euaa172
27. Dimopoulos AC, Nikolaidou M, Caballero FF, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Med Res Methodol*. 2018;18(1):1-11. doi:10.1186/s12874-018-0644-1
28. Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol*. 2012;176(6):473-481. doi:10.1093/aje/kws207
29. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic Progn Res*. 2020;4(1):4-9. doi:10.1186/s41512-020-00077-0
30. Smeets HM, De Wit NJ, Hoes AW. Routine health insurance data for scientific research: Potential and limitations of the Agis Health Database. *J Clin Epidemiol*. 2011;64(4):424-430. doi:10.1016/j.jclinepi.2010.04.023
31. Smeets HM, Kortekaas MF, Rutten FH, et al. Routine primary care data for scientific research, quality of care programs and educational purposes: The Julius General Practitioners' Network (JGPN). *BMC Health Serv Res*. 2018;18(1):1-9. doi:10.1186/s12913-018-3528-5

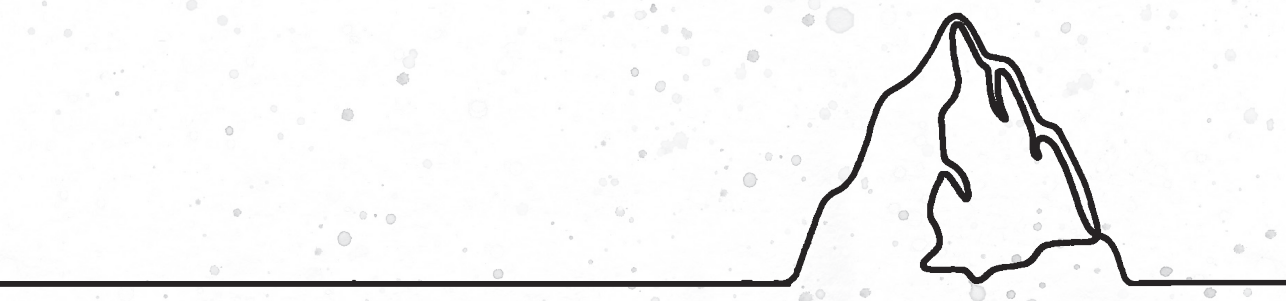
32. Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J Am Med Informatics Assoc*. 2016;23(6):1143-1149. doi:10.1093/jamia/ocw021
33. Boulton C, Wilkinson JM. Use of public datasets in the examination of multimorbidity: Opportunities and challenges. *Mech Ageing Dev*. 2020;190(xxxx):111310. doi:10.1016/j.mad.2020.111310
34. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59(10):1087-1091. doi:10.1016/j.jclinepi.2006.01.014
35. Nijman SWJ, Hoogland J, Groenhof TKJ, et al. Real-time imputation of missing predictor values in clinical practice. *Eur Hear J - Digit Heal*. 2021;2(1):154-164. doi:10.1093/ehjdh/ztaa016
36. Perez M V., Mahaffey KW, Hedlin H, et al. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N Engl J Med*. 2019;381(20):1909-1917. doi:10.1056/nejmoa1901183
37. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson*. 2018;20(65):1-12. doi:10.1186/s12968-018-0471-x
38. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394(10201):861-867. doi:10.1016/S0140-6736(19)31721-0
39. Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering*. 2020;6(3):291-301. doi:10.1016/j.eng.2019.08.015
40. Herzlinger RE. Why innovation in health care is so hard. *Harv Bus Rev*. 2006;(May).
41. Dixon-Woods M, Amalberti R, Goodman S, Bergman B, Glasziou P. Problems and promises of innovation: Why healthcare needs to rethink its love/hate relationship with the new. *BMJ Qual Saf*. 2011;20(SUPPL. 1):47-51. doi:10.1136/bmjqs.2010.046227
42. Dhindsa K, Bhandari M, Sonnadara RR. What's holding up the big data revolution in healthcare? *BMJ*. 2018;363(December):1-2. doi:10.1136/bmj.k5357
43. Peloquin D, Dimaio M, Bierer B, Barnes M. Disruptive and avoidable : GDPR challenges to secondary research uses of data. *Eur J Hum Genet*. Published online 2020:697-705. doi:10.1038/s41431-020-0596-x
44. Bahls T, Pung J, Heinemann S, et al. Designing and piloting a generic research architecture and workflows to unlock German primary care data for secondary use. *J Transl Med*. Published online 2020:1-10. doi:10.1186/s12967-020-02547-x
45. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol*. 2020;145(2):463-469. doi:10.1016/j.jaci.2019.12.897
46. Spasic I, Nenadic G. Clinical text data in machine learning: Systematic review. *JMIR Med Informatics*. 2020;8(3). doi:10.2196/17984
47. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*. 2020;323(4):305-306. doi:10.1001/jama.2019.20866
48. Federer LM, Belter CW, Joubert DJ, et al. Data sharing in PLOS ONE : An analysis of Data Availability Statements. *PloS One*. 2018;13(5): e0194768. doi: 10.1371/journal.pone.0194768
49. Bender D, Sartipi K. HL7 FHIR: An agile and RESTful approach to healthcare information exchange. *Proc CBMS 2013 - 26th IEEE Int Symp Comput Med Syst*. Pub-

lished online 2013:326-331. doi:10.1109/CBMS.2013.6627810
50. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. *J Biomed Inform*. 2019;94(May):103188. doi:10.1016/j.jbi.2019.103188
51. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. *J Biomed Inform*. 2013;46(1):87-96. doi:10.1016/j.jbi.2012.09.006
52. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digit Med*. 2020;81. doi:10.1038/s41746-020-0288-5
53. Hnatkova K, Smetana P, Toman O, Schmidt G, Malik M. Sex and race differences in QRS duration. *Europace*. 2016;18(12):1842-1849. doi:10.1093/europace/euw065
54. van der Ende MY, Siland JE, Snieder H, van der Harst P, Rienstra M. Population-based values and abnormalities of the electrocardiogram in the general Dutch population: The LifeLines Cohort Study. *Clin Cardiol*. 2017;40(10):865-872. doi:10.1002/clc.22737
55. Jones J, Srodulski ZM, Romisher S. The aging electrocardiogram. *Am J Emerg Med*. 1990;8(3):240-245. doi:10.1016/0735-6757(90)90331-S
56. Rijnbeek PR, Van Herpen G, Bots ML, et al. Normal values of the electrocardiogram for ages 16-90 years. *J Electrocardiol*. 2014;47(6):914-921. doi:10.1016/j.jelectrocard.2014.07.022
57. Noseworthy PA, Attia ZI, Brewer LPC, et al. Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis. *Circ Arrhythmia Electrophysiol*. 2020;(March):208-214. doi:10.1161/CIRCEP.119.007988
58. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med*. 2021;13(586). doi:10.1126/scitranslmed.abb1655
59. Gundersen OE, Gil Y, Aha DW. On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI Mag*. 2018;39(3):56-68. doi:10.1609/aimag.v39i3.2816
60. McKinney SM, Sieniek M, Godbole V, et al. Addendum: International evaluation of an AI system for breast cancer screening. *Nature*. 2020;586(7829):E19. doi:10.1038/s41586-020-2679-9
61. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6
62. Tat E, Bhatt DL, Rabbat MG. Comment Addressing bias : artificial intelligence in cardiovascular medicine. *Lancet Digit Heal*. 2020;2(12):e635-e636. doi:10.1016/S2589-7500(20)30249-1
63. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
64. Kentner AC, Grace SL. Between mind and heart: Sex-based cognitive bias in cardiovascular disease treatment. *Front Neuroendocrinol*. 2017;45:18-24. doi:10.1016/j.yfrne.2017.02.002
65. Leening MJG, Elias-Smale SE, Felix JF, et al. Unrecognised myocardial infarction and long-term risk of heart failure in the elderly: The Rotterdam study. *Heart*. 2010;96(18):1458-1462. doi:10.1136/hrt.2009.191742
66. Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net Reclassification Improvement: Computation, Interpretation and Controversies. *Ann Intern Med*. 2014;160(2):122-131. doi:10.7326/M13-1522
67. Lee HC, Park JS, Choe JC, et al. Prediction

of 1-Year Mortality from Acute Myocardial Infarction Using Machine Learning. *Am J Cardiol*. 2020;133:23-31. doi:10.1016/j.amjcard.2020.07.048
68. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*. 2021;31(6):3797-3804. doi:10.1007/s00330-021-07892-z
69. Fryback DG, Thornbury JR. The Efficacy of Diagnostic Imaging. *Med Decis Mak*. 1991;11(2):88-94. doi:10.1177/0272989X9101100203
70. Topol EJ. Welcoming new guidelines for AI clinical research. *Nat Med*. 2020;26(9):1318-1320. doi:10.1038/s41591-020-1042-x
71. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*. 2017;284(2):574-582. doi:10.1148/radiol.2017162326
72. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black Box" Medicine? *Ann Intern Med*. 2020;172(1):59-61. doi:10.7326/M19-2548
73. Price II WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA - J Am Med Assoc*. 2019;322(18):1765-1766. doi:10.1001/jama.2019.4914
74. Ploug T, Holm S. The four dimensions of contestable AI diagnostics- A patient-centric approach to explainable AI. *Artif Intell Med*. 2020;107(January):101901. doi:10.1016/j.artmed.2020.101901
75. Bjerring JC, Busch J. Artificial Intelligence and Patient-Centered Decision-Making. *Philos Technol*. 2021;34:349-371. doi:https://doi.org/10.1007/s13347-019-00391-6
76. van Leeuwen KG, Siegersma KR. Van data tot patiënt: Ontwikkeling kunstmatige intelligentie vergt lange adem. *Med Contact* (Bussum). 2021;(April):7-9.
77. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care; Will the value match the hype? *JAMA - J Am Med Assoc*. 2019;321(23):2281-2282. doi:10.1001/jama.2019.4914

Appendix

English summary
Nederlandse samenvatting
List of contributing authors
List of publications
PhD portfolio
Dankwoord
Curriculum Vitae



SUMMARY

The aim of this thesis "Moving from traditional methods towards artificial intelligence in cardiovascular research with regular care data" was to investigate the use of different research methods, varying from traditional statistics towards artificial intelligence, on regular care data. A specific focus in this thesis was the presentation of cardiovascular disease in women, a subgroup that is structurally underrepresented in clinical research. The regular care database used for the larger part of chapters of this thesis was described in **Chapter 2**. This database encompasses all patients suspected of cardiovascular disease who visited one of the Cardiology Centers of the Netherlands (CCN) between 2007 and 2018 (n = 109,151). Long-term follow-up was available after linkage with the personal registry data of Statistics Netherlands (CBS). Strengths of the CCN database are, first, the large population included in the database. This population reflects the patients that are currently seen in regular cardiac care and thus includes groups that are structurally underrepresented or excluded in clinical trials, e.g. women, elderly patients and the patients with comorbidities. Second, the database includes a large number of different, and in some patients longitudinal, measurements. Third, regular care databases reflect contemporary medical practice. This allows for comparisons between regular care and guideline recommendations. These perspectives allow for debate on discrepancies between guideline-recommended care and daily clinical practice. Nonetheless, the use of regular care data has limitations. First, data quality does not meet standards of clinical trials, as data has been collected for care purposes and not for research. Therefore, data collection and follow-up were not uniform across patients. This also leads to missing variables across patients as not all diagnostics are performed in each patient. Second, not all data is structurally included in the database and free text fields are common. Therefore, thorough data cleaning is required, that might include text mining approaches.

In the first part we used the database introduced in Chapter 2 to evaluate the current shift in cardiac practice to replace traditional electrocardiography (ECG) stress testing by cardiac computed tomography (CT) for calcium scoring and angiography to diagnose obstructive coronary heart disease in symptomatic patients that are suspected of chronic coronary syndrome; the so-called CT-first strategy. **Chapter 3** used traditional survival analysis to evaluate the effects of a CT-first strategy in patients with chest pain with regular care data. Methods to make a regular care database fit for research purposes were used in this chapter. This included multiple imputation for chained equations to account for missing values, and propensity score matching, to make similar groups between patients with chest pain that had a CT-first strategy and patients that did not have a cardiac CT after their chest pain consult. This study showed that a CT-first strategy reduces all-cause mortality, but not cardiovascular mortality. In **Chapter 4** the results of the cardiac CT were used to predict mortality in patients with chest pain. As previous studies did

not perform a sex-stratified analysis, this study specifically focussed on the prognostic value of the coronary artery calcium score and race-, age- and sex-specific percentiles of the coronary artery calcium score in the sexes. Both measures predict mortality equally well, in men and women. Hence, there is no need for personalized percentiles based on sex, age and race for risk stratification. Furthermore, evaluation of the addition of degree of stenosis to coronary artery calcium score for risk stratification, showed increased discrimination, although non-significant, in women compared to men. Nonetheless, this is clinically relevant finding, as it supports a sex-specific view on coronary artery disease as women more frequently present with non-calcified plaques than men. **Chapter 5A** and **5B** evaluated the New York Heart Association classification (NYHA) class for risk assessment in all patients visiting one of the CCN centers, despite of their initial complaints. Patients were selected with complaints of chest pain, dyspnoea and fatigue and a reported NYHA classification. This showed that the use and value of the NYHA classification can be extended beyond complaints related to heart failure. **Chapter 5A** is a short report to investigate the value of the NYHA class in women, as most studies primarily included men suspected of heart failure. We showed that a higher NYHA class is associated with an increased mortality risk in women with chest pain, dyspnoea and fatigue. A similar trend was shown in men, although hazard ratios were higher in men compared to women. As NYHA-class is a subjective measure of the experienced daily disability by the patient, **Chapter 5B** focussed on the explanation of the NYHA-class with objective measures in both men and women. For that purpose, a mediation analysis was performed to study the causality of the stress ECG variables on the relation between NYHA classification and all-cause mortality in subgroup of patients. Proportional workload (i.e. maximum work load during exercise as a proportion of the predicted work load) appeared to be the largest mediator in the association between NYHA classification and mortality in men and women, but the majority of the association remains unexplained.

In the second part, we described the use of different methodologies that are classified as artificial intelligence (AI) for cardiovascular research with regular care data. This part of the thesis used different data sources for machine learning modelling; free text, cardiovascular imaging modalities, regular care data extracted from electronic health records and raw ECG data. In **Chapter 6** a pipeline was developed for the identification of adverse drug reactions in clinical notes of the cardiologist. The pipeline used word embedding models, that were trained on all the clinical notes available in the regular care dataset. Although the performance did not reach up to specific models that were developed for the English language, overall performance and set-up of the pipeline was good. Furthermore, the pipeline facilitated interpretation and can be easily tuned for other applications. **Chapter 7** gives a narrative overview of the available opportunities that apply AI to cardiovascular imaging. Currently, AI applications are developed for car-

diac magnetic resonance imaging, cardiac computed tomography, echocardiography and nuclear cardiac imaging. AI will impact all steps in the cardiovascular imaging chain, i.e. automatic selection of imaging protocol, automated segmentation of cardiovascular structures, and risk stratification based on cardiac imaging. Although these models showed their value in a research setting, clinical implementation is limited, due to among other reasons, a lack of validation and certification. **Chapter 8** described the use of regular care data derived from the electronic health record to study the performance and improvement of the pre-test probability of coronary artery disease in patients with chest pain or dyspnoea. The pre-test probability is a risk stratification tool, based on sex, age and type of complaints, and is recommended by the 2019 European Society of Cardiology guidelines for chronic coronary syndromes to refer patients for non-invasive imaging. We showed that this tool misclassified approximately 20% of women with a diagnosis of coronary artery disease, as set by the treating cardiologist. This means that the pre-test probability predicted that these women had less than 5% risk of CAD, and therefore these women would have not been referred for further diagnostic imaging test. The prediction of the presence of CAD significantly improved when all available data in the electronic health records was used and a model was constructed using a Lasso logistic regression. However, more sophisticated gradient boosting models did not improve the classification any further. In **Chapter 9** a large number of ECGs made in regular clinical care were used to develop a deep neural network that classified sex based on the raw ECG data. This model was validated in two external validation datasets from population-based studies and showed accurate performance for the prediction of sex. Moreover, it was shown that individuals that were misclassified based on their ECG had worse survival than their correctly classified biological peers. Mediation analysis revealed an undiscovered relation between a shortened QRS duration and increased mortality risk in both men and women. This chapter emphasizes the importance of sex-stratified research and implementation of sex in study design.

Chapter 10 reflects on the use of traditional and AI methods for cardiovascular research with regular care data. AI has the potential to automatically analyse large amounts of data and to pick up associations that have not been discovered with traditional methods. AI is specifically useful when data is unstructured, i.e. free text fields, cardiovascular imaging or ECG signals, as these types of data cannot be (automatically) analysed with traditional methods. Nonetheless, the clinical implementation of AI models is still limited. Multiple causes can be appointed for this delay in implementation; a lack of external validation studies, reduced generalizability of AI models, reproducibility and replicability issues and a lack of knowledge of clinical efficacy.

SAMENVATTING

Het doel van dit proefschrift, getiteld "Moving from traditional methods towards artificial intelligence in cardiovascular research with regular care data", is om het gebruik van verschillende onderzoeksmethoden, variërend van traditionele statistiek tot kunstmatige intelligentie, te onderzoeken. Hiervoor is gebruik gemaakt van reguliere zorgdata uit het patiëntendossier. Deze data is verzameld in de dagelijkse praktijk van de cardioloog. Specifieke aandacht is er in dit proefschrift voor verschillen tussen mannen en vrouwen met hart- en vaatziekten. Vrouwen zijn namelijk een subgroep die structureel ondervertegenwoordigd is in het klinisch-wetenschappelijk onderzoek. Het proefschrift start met een algemene introductie en wordt vervolgd met een hoofdstuk waarin de database beschreven wordt die gebruikt is voor het onderzoek gepresenteerd in dit proefschrift (**Hoofdstuk 2**). Deze database bevat alle patiënten met een vermoeden van hart- en vaatziekten die tussen 2007 en 2018 één van de klinieken van Cardiologie Centra Nederland (CCN) hebben bezocht (n=109,151). Na koppeling van de CCN database met de databases van het Centraal Bureau voor de Statistiek (CBS) is langdurige follow-up beschikbaar van deze patiënten. Het gebruik van de CCN database heeft een aantal sterke punten wat deze database geschikt maakt voor onderzoek. Ten eerste bevat deze database een zeer grote populatie met mogelijk hart- en vaatziekten. Deze populatie weerspiegelt de patiënten die in de hedendaagse praktijk van de cardioloog komen en bevat dus ook groepen die structureel ondervertegenwoordigd of uitgesloten zijn in klinische onderzoeken, bijvoorbeeld vrouwen, oudere patiënten en patiënten met meerdere onderliggende aandoeningen, zoals diabetes en hypertensie. Ten tweede bevat de database een groot aantal verschillende metingen. Bij een aantal patiënten is dezelfde meting meerdere keren gedaan tijdens vervolgbezoeken. Ten derde weerspiegelen reguliere zorgdatabases de hedendaagse medische praktijk. Dit maakt het mogelijk om een vergelijking te maken tussen reguliere zorg die gegeven wordt en de voorgeschreven richtlijnen. Dit perspectief geeft inzicht in de discrepanties tussen richtlijnen en de dagelijkse klinische praktijk. Het gebruik van reguliere zorgdata voor onderzoek kent echter ook beperkingen. Ten eerste voldoet de kwaliteit van de gegevens niet aan de normen van klinisch-wetenschappelijk onderzoek, omdat gegevens zijn verzameld voor zorgdoeleinden en niet voor onderzoek. Gegevensverzameling en follow-up zijn dus ook niet uniform voor alle patiënten. Dit leidt tot ontbrekende variabelen bij patiënten, aangezien niet alle diagnostiek bij elke patiënt wordt uitgevoerd. Ten tweede worden niet alle gegevens structureel in de database opgenomen en zijn vrije tekstvelden gebruikelijk. Daarom is eerst een grondige herstructurering en filtering van de data nodig, bijvoorbeeld met behulp van automatische analyse van vrije tekstvelden.

Het eerste deel van dit proefschrift gebruikt de database, geïntroduceerd in Hoofdstuk 2, om de verschuiving van functionele naar beeldvormende diagnostiek te evalueren

voor patiënten met pijn op de borst. In de klinische praktijk worden steeds minder inspanningstesten met registratie van het electrocardiogram (ECG) afgenomen. In plaats daarvan wordt de keuze gemaakt voor een diagnostische cardiale computed tomography-scan (CT), waarbij de hoeveelheid kalk en vernauwingen in de kransslagaders beoordeeld worden; dit is de zogenaamde CT-first strategie. Op deze manier wordt de diagnose 'chronisch coronair syndroom' vastgesteld. **Hoofdstuk 3** onderzocht verschillen in overleving tussen patiënten met pijn op de borst die een CT gehad hebben tijdens hun diagnostische traject en patiënten zonder een CT. Met behulp van "multiple imputation for chained equations" (MICE) zijn missende variabelen bij patiënten ingevuld en "propensity score matching" is gebruikt om vergelijkbare groepen te maken tussen de patiënten mét en zonder CT. Deze studie toont aan dat bij het toepassen van de CT-first strategie het risico op overlijden van patiënten met pijn op de borst significant lager is. Daarentegen is het verschil in overlijden door een cardiovasculaire oorzaak niet verschillend tussen de patiënten met en zonder een CT-first strategie. In **Hoofdstuk 4** werden de resultaten van de cardiale CT gebruikt om overleving bij patiënten met pijn op de borst te voorspellen. In eerdere studies zijn geen analyses gedaan voor mannen en vrouwen apart. De gepresenteerde studie richt zich specifiek op het vergelijken van de voorspellende waarde van de algemene kalkscore en de etniciteits-, leeftijds- en geslachtsspecifieke percentielen (MESA-percentielen) van de kalkscore bij mannen en vrouwen. De studie laat zien dat beide representaties van de kalkscore sterfte even goed voorspellen, zowel bij mannen als bij vrouwen. Er is dus geen noodzaak voor het gebruik van gepersonaliseerde percentielen op basis van geslacht, leeftijd en etniciteit voor risicostratificatie bij patiënten met mogelijk chronisch coronairlijden. Bij vrouwen verbetert de voorspellende waarde van de kalkscore na toevoeging van de mate van vernauwing. Dit is een klinisch zeer relevante bevinding, ondanks dat deze niet statistisch significant is. Het ondersteunt de geslachtsspecifieke kijk op coronairlijden en bevestigt de hypothese dat vrouwen vaker niet-verkalkte plaques. Deze worden niet in de kalkscore weerspiegeld.
In **Hoofdstuk 5A** en **5B** werd de waarde van de New York Heart Association (NYHA) functionele classificatie onderzocht voor risicostratificatie bij patiënten die één van de klinieken van CCN bezocht hadden. Patiënten met een gerapporteerde NYHA classificatie zijn hiervoor geselecteerd uit de database. Deze patiënten presenteerden zich met een verscheidenheid aan klachten; pijn op de borst, dyspnoe en vermoeidheid. **Hoofdstuk 5A** is een korte rapportage om de waarde van de NYHA classificatie bij vrouwen te onderzoeken, aangezien de meeste studies voornamelijk mannen, includeren. Dit hoofdstuk laat zien dat een hogere NYHA classificatie geassocieerd is met een verhoogd risico op overlijden bij vrouwen met pijn op de borst, dyspneu en vermoeidheid. Een vergelijkbare, sterkere associatie is waargenomen bij mannen. Aangezien de NYHA classificatie een subjectieve maat is die omschrijft hoe de patiënt het dagelijkse leven ervaart met

invaliderende klachten, concentreert **Hoofdstuk 5B** zich op het verklaren van de NYHA classificatie met objectieve metingen bij mannen en vrouwen. Hiervoor is een mediatieanalyse uitgevoerd om te beoordelen in hoeverre metingen, uitgevoerd tijdens het inspannings-ECG, een effect hebben op de relatie tussen de NYHA classificatie en het risico op overlijden. Proportionele inspanningsbelasting (d.w.z. de maximale inspanningsbelasting tijdens het inspannings-ECG als percentage van de vooraf voorspelde inspanningsbelasting op basis van geslacht, leeftijd en lengte) heeft het grootste effect op de associatie tussen de NYHA classificatie en mortaliteit van zowel mannen als vrouwen, maar het grootste deel van de relatie blijft onverklaard.

In de tweede helft van dit proefschrift is het gebruik van kunstmatige intelligentie (artificial intelligence, AI) voor cardiovasculair onderzoek met reguliere zorgdata onderzocht. Dit deel van het proefschrift gebruikt verschillende soorten data voor analyse; vrije tekstvelden, cardiovasculaire beeldvorming, data geregistreerd in het elektronisch patiëntendossier en ECG data. In **Hoofdstuk 6** is een digitale workflow ontwikkeld voor de identificatie en extractie van bijwerkingen in klinische notities van de cardioloog. Deze workflow maakt gebruikt van word-embedding modellen, gebaseerd op alle klinische notities die opgenomen zijn in de CCN database. Ondanks dat de prestaties van de workflow niet overeenkomen met modellen die specifiek ontwikkeld zijn voor Engelse klinische notities, is de algehele uitvoering en opzet goed. Bovendien kan de workflow gemakkelijk aangepast worden naar andere toepassingen van automatische extractie uit vrije tekst. **Hoofdstuk 7** geeft een overzicht van verscheidene toepassingen van kunstmatige intelligentie in de cardiovasculaire beeldvorming. AI-toepassingen worden op dit moment ontwikkeld voor beeldvorming van het hart met verschillende modaliteiten: magnetische resonantie (magnetic resonance imaging, MRI), CT, ultrasound en nucleaire beeldvorming. Kunstmatige intelligentie heeft invloed op alle stappen in de keten van beeldvorming, d.w.z. automatische selectie van het protocol, geautomatiseerde segmentatie en interpretatie van cardiovasculaire structuren en risicostratificatie op basis van cardiale beeldvorming. Hoewel deze modellen hun waarde hebben bewezen in het onderzoek, is de klinische implementatie beperkt, onder meer door gebrek aan validatie en certificering. In **Hoofdstuk 8** is het gebruik van gegevens uit het elektronisch patiëntendossier bestudeerd, met als doel het verbeteren van de pre-test waarschijnlijkheid (pre-test probability, PTP) op coronairlijden bij patiënten met pijn op de borst of dyspneu. De PTP is een instrument voor het inschatten van risico op coronairlijden op basis van geslacht, leeftijd en type klachten en wordt gebruikt om patiënten door te verwijzen voor niet-invasieve beeldvorming. In ongeveer 20% van de vrouwen met coronairlijden wordt het risico te laag ingeschat (<5% risico op coronairlijden op basis van de PTP). Dit heeft als gevolg dat deze groep vrouwen niet doorverwezen wordt voor verdere diagnostische beeldvorming van het hart. De risicoclassificatie verbetert significant wan-

neer alle beschikbare gegevens in het elektronische patiëntendossier gebruikt worden en een model wordt getraind met behulp van Lasso logistische regressie. Meer geavanceerde modellen verbeteren de classificatie echter niet meer dan logistische regressie. In **Hoofdstuk 9** is een deep neural network (DNN) ontwikkeld dat op basis van het ECG geslacht kan classificeren van de patiënt. Dit model is gevalideerd in twee andere populaties en kan ook in deze datasets het geslacht correct classificeren in het grootste deel van de ECGs. Bovendien is aangetoond dat individuen die verkeerd zijn geclassificeerd op basis van hun ECG een slechtere overleving hebben dan hun correct geclassificeerde biologische leeftijdsgenoten. Mediatieanalyse laat een relatie zien tussen een verkorte QRS-duur op het ECG en een verhoogd risico op overlijden bij zowel mannen als vrouwen. Deze relatie is nog niet eerder beschreven. Dit hoofdstuk benadrukt het belang van geslachtsspecifiek onderzoek en de implementatie van geslacht in de onderzoeksopzet. **Hoofdstuk 10** reflecteert op het gebruik van traditionele methodes en kunstmatige intelligentie in cardiovasculair onderzoek met reguliere zorgdata. AI heeft de potentie om automatisch grote hoeveelheden data te analyseren en relaties te identificeren die met traditionele methodes niet eerder onderzocht zijn. Kunstmatige intelligentie is met name toepasbaar wanneer gegevens ongestructureerd zijn, d.w.z. vrije tekstvelden, cardiovasculaire beeldvorming of ECG-signalen, aangezien dit soort gegevens niet (automatisch) kunnen worden geanalyseerd met traditionele methodes. Desalniettemin is de klinische implementatie van AI-modellen nog beperkt. Voor deze discrepantie tussen potentie en implementatie zijn meerdere oorzaken aan te wijzen; een gebrek aan externe validatie, verminderde generaliseerbaarheid van AI-modellen, problemen met reproductie en replicatie van resultaten en een gebrek aan inzicht in werking van het model.

LIST OF CONTRIBUTING AUTHORS

Yolande Appelman, MD PhD
Department of Cardiology, Amsterdam University Medical Centres, location VU University Amsterdam, Amsterdam, The Netherlands

Prof. Folkert W. Asselbergs, MD
Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Sophie H. Bots, PhD
Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Prof. Michiel L. Bots, MD
Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Willemijn J. op den Brouw, MSc.
Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Prof. Derek P. Chew, MD
Department of Cardiovascular Medicine, Flinders Medical Centre, Bedford Park, SA, Australia
South Australian Health and Medical Research Institute, Adelaide, SA, Australia

Ruben Coronel, MD PhD
Heart Center, Department of Experimental Cardiology, location AMC, Amsterdam University Medical Centres, Amsterdam, The Netherlands

Ernest Diez-Benavente, PhD
Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Prof. Pieter Doevendans, MD
Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
Netherlands Heart Institute, Utrecht, The Netherlands

Anouk M. Eikendal, MD PhD
Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

René van Es, PhD
Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Maxime Evers, MSc.
Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Floor Groepenhoff, MD, PhD
Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
Central Diagnostic Laboratory, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Prof. Pim van der Harst, MD
Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Rutger J. Hassink, MD PhD
Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Prof. Leonard Hofstra, MD
Department of Cardiology, Amsterdam University Medical Centres, location VU University Amsterdam, Amsterdam, The Netherlands
Cardiology Centers of the Netherlands, Utrecht, The Netherlands

Prof. Tim Leiner, MD
Department of Radiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Prof. David A. Leon
Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom
International Laboratory for Population and Health, National Research University, Higher School of Economics, Moscow, Russian Federation
Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway

Rutger R. van de Leur, MD
Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
Netherlands Heart Institute, Utrecht, The Netherlands

Prof. Jagat Narula, MD
Icahn School of Medicine at Mount Sinai, New York, New York, USA

N. Charlotte Onland-Moret, PhD
Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Liesbeth Rozendaal, MD
Julius Gezondheidscentrum Parkwijk, Utrecht, the Netherlands

Prof. ir. Hester M. den Ruijter
Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

G. Aernout Somsen, MD PhD
Cardiology Centers of the Netherlands, Utrecht, The Netherlands

Prof. Marco Spruit
Department of Public Health and Primary Care (PHEG), Leiden University Medical Center (LUMC), Leiden University, Leiden, The Netherlands
Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, The Netherlands

Igor I. Tulevski, MD PhD
Cardiology Centers of the Netherlands, Utrecht, The Netherlands

Johan W. Verjans, MD PhD
Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia
Department of Cardiology, Royal Adelaide Hospital, Adelaide, SA, Australia
South Australian Health and Medical Research Institute, Adelaide, SA, Australia

LIST OF PUBLICATIONS

Journal Articles

Siegersma, K. R., Leiner, T., Chew, D. P., Appelman, Y., Hofstra, L., & Verjans, J. W. (2019). Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist. *Netherlands heart journal* , 27(9), 403–413. https://doi.org/10.1007/s12471-019-01311-1

Siegersma, K. R.[*], Groepenhoff, F.[*], Onland-Moret, N. C., Tulevski, I. I., Hofstra, L., Somsen, G. A.[#], & Den Ruijter, H. M.[#] (2021). New York Heart Association class is strongly associated with mortality beyond heart failure in symptomatic women. *European heart journal. Quality of care & clinical outcomes*, 7(2), 214–215. https://doi.org/10.1093/ehjqcco/qcaa091

Bots, S. H.[*], Siegersma, K. R.[*], Onland-Moret, N. C., Asselbergs, F. W., Somsen, G. A., Tulevski, I. I., den Ruijter, H. M., & Hofstra, L. (2021). Routine clinical care data from thirteen cardiac outpatient clinics: design of the Cardiology Centers of the Netherlands (CCN) database. *BMC cardiovascular disorders*, 21(1), 287. https://doi.org/10.1186/s12872-021-02020-7

Siegersma, K. R., Onland-Moret, N. C., Appelman, Y., van der Harst, P., Tulevski, I. I., Somsen, G. A., Narula, J., den Ruijter, H. M., & Hofstra, L. (2021). Outcomes in patients with a first episode of chest pain undergoing early coronary CT imaging. *Heart,* 108, 1361-1368. https://doi.org/10.1136/heartjnl-2021-319747

Siegersma, K. R., Evers, M., Bots, S. H., Groepenhoff, F., Appelman, Y., Hofstra, L., Tulevski, I. I., Somsen, G. A., den Ruijter, H. M., Spruit, M.[#], & Onland-Moret, N. C.[#] (2022). Development of a Pipeline for Adverse Drug Reaction Identification in Clinical Notes: Word Embedding Models and String Matching. *JMIR medical informatics*, 10(1), e31063. https://doi.org/10.2196/31063

Siegersma, K. R.[*], van de Leur, R. R.[*], Onland-Moret, N. C., Leon, D. A., Diez-Benavente, E., Rozendaal, L., et al., Bots, M. L., Coronel, R., Appelman, Y., Hofstra, L., van der Harst, P., Doevendans, P. A., Hassink, R. J., den Ruijter, H. M.[#], van Es, R.[#] (2022). Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. *European heart journal. Digital Health*, 3(2), 245-254. https://doi.org/10.1093/ehjdh/ztac010

Siegersma, K. R.[*], Stens, N. A.[*], Groepenhoff, F., Appelman, Y., Tulevski, I. I., Hofstra, L., den Ruijter, H.M., Somsen, G. A.[#], Onland-Moret, N.C.[#] (2022). Sex differences in the relationship between New York Heart Association functional classification and survival in cardiovascular disease patients: A mediation analysis of exercise capacity with regular care data. *Reviews of Cardiovascular Medicine,* 23(8), 278. https://doi.org/10.31083/j.rcm2308278

[*#]: Authors contributed equally

Submitted

Siegersma, K. R.[*], Groepenhoff, F.[*], Eikendal, A. L. M., op den Brouw W. J., Leiner T., Appelman Y., Tulevski I. I., Somsen G. A., Onland-Moret N. C., Hofstra, L.2,5[#], den Ruijter H. M.[#] (2022). Coronary calcification measures predict mortality in symptomatic women and men. *Submitted*

In preparation

Siegersma, K. R., Hofstra, L., Appelman, Y., Benjamins, J. W., van der Harst, P., Tuleski, I. I., den Ruijter, H. M., Somsen, G. A.*, Onland-Moret, N. C.* (2022), Improving the classification of women at high risk of coronary artery disease with logistic regression and gradient boosting using a regular care database. *In preparation*

Conference abstracts and presentations

Siegersma, K. R., Groepenhoff, F., van Es, B., Blasse, E., Bots, S. H., Peeters, T., de Groot, M., van Solinge, W., Höfer, I., Tauber, T., den Ruijter, H. M., Haitjema, S., (2019, October 5-6). *ARGUS Hackathon: Development of a machine learning algorithm to exclude coronary macro- and microvascular disease in a clinical care dataset* [conference presentation]. ESC Digital Summit. Tallinn, Estonia.

Siegersma, K.R., Appelman, Y., den Ruijter, H. M., Tulevski, I. I., Somsen, G. A., Hofstra, L., (2019, October 5-6). *Development of a decision support tool to predict outcome of coronary CT calcium score and angiography: Study Design and Data Preparation* [conference presentation]. ESC Digital Summit. Tallinn, Estonia.

Siegersma, K.R., van de Leur, R. R., Onland-Moret, N.C., van Es, R. & den Ruijter, H. M., (2020, August 27-30). *Misclassification of sex by deep neural networks reveals novel ECG characteristics that explain a higher risk of mortality in women and in men* [conference presentation]. ESC Congress 2021 - The Digital Experience.

PHD PORTFOLIO

PhD candidate: K.R. Siegersma
Department: Cardiologie, Amsterdam UMC, location VUmc
Graduate school: Amsterdam Cardiovascular Sciences
PhD period: November 1, 2018 - October 31, 2021
Promotor: Prof. dr. L. Hofstra, Prof. dr. ir. H.M. den Ruijter
Copromotors: dr. N.C. Onland-Moret, dr. Y. Appelman

Teaching activities	Year(s)	ECTS
Lectures and mentoring		
The use of artificial intelligence in cardiovascular research and care	2019	0.25
PhD project presentation at SUMMA symposium	2019	0.14
ESCR webinar: AI in cardiovascular imaging - to be or not to be?	2020	0.50
Project mentor Vascular Biology (Biomedical Sciences)	2020, 2021	0.50
Supervision of internships		
Portfolio assignment Applied Data Science	2019	0.50
Bachelor project Medische Informatiekunde	2020	2.14
Master project Business Informatics	2020	2.57
Literature review Biomedical Sciences	2021	0.50
Master project Biomedical Sciences and Epidemiology	2021	3.00

Personal and professional development	Year(s)	ECTS
Participant Hackathon Hartstichting, ABN AMRO, UMC Utrecht	2019	0.75
Be your best selfie	2019	0.14
Grant writing Horizon2020	2020	2.00
Career orientation and coaching	2020	0.25
Medezeggenschap brainstorm CNV Jongeren	2020	0.50
Circulatory Health Coffee (organization)	2020, 2021	2.00
Young Science in Transition	2020, 2021	0.75

Training activities	Year(s)	ECTS
Courses & Workshops		
Patient Participation	2019	
Basiscursus Regelgeving en Organisatie voor Klinisch Onderzoekers (BROK)	2019	1.50
Writing a Data Management Plan	2019	1.00
Cardiovascular course NHS: Cardiac Function & Adaptation	2019	2.00
Reading group Deep Learning	2019-2020	2.43
Reading group Natural Language Processing	2019-2020	1.71
Grant writing for junior researchers	2020	0.14
Zakelijk tekenen	2020	0.14
Research Integrity	2021	2.00
Scientific writing in English	2021	1.50
Special interest group Applied Data Science	2019-2021	1.50
Symposia, Conferences & Seminars		
PhD retreat Amsterdam Cardiovascular Science (presentation)	2019	1.00
Technical Innovations in Medicine Conference (organization)	2019	2.00
Young@Heart: Sell your science (presentation)	2019	0.25
Third Translational Cardiovascular Research Meeting	2019	0.25
Scientific session E-health Netherlands Heart Institute	2019	0.10
NVvTG Wetenschapsavond: Artificial Intelligence (presentation)	2019	0.25
ESC Digital Summit (presentation)	2019	2.00
Young@Heart: How to conquer the world with your PhD? (organization)	2019	1.00
Fourth Translational Cardiovascular Research Meeting	2020	0.10
European Society of Cardiology congress	2020	1.00
Fifth Translational Cardiovascular Research Meeting	2021	0.25
Circulatory Health Summer Meeting (organization)	2021	0.50
European Society of Cardiology congress (presentation)	2021	2.00
Highlights session European Society of Cardiology congress (presentation)	2021	0.50
Research presentations and brainstorms Cardiology department UMC Utrecht	2019, 2020, 2021	2.00
PhD afternoons Amsterdam Cardiovascular Sciences	2019, 2020, 2021	1.50

DANKWOORD

"Ik heb het nog nooit gedaan, dus ik denk dat ik het wel kan." – *Pippi Langkous*

Elk nieuw avontuur begint voor mij vanuit deze quote van Pippi Langkous (mijn jeugdheld), anders zou ik geen tijd aan een nieuw avontuur besteden. Ruim 3,5 jaar geleden begon ik aan mijn PhD-avontuur met dit proefschrift als resultaat. Ik dacht dat ik het wel kon, maar zonder een groot aantal bijzondere mensen had ik het niet gekund!

Beste **Hester**, **prof. dr. ir. Den Ruijter**, wat ben ik blij dat ik in jouw onderzoeksgroep terecht ben gekomen. In het begin was het schipperen tussen mijn rol in de VU als PhD-student en mijn praktische werkzaamheden die toch voornamelijk in het UMC Utrecht waren, maar ik ben heel blij dat ik (en eigenlijk wij samen) daar een weg in gevonden hebben. Ik bewonder je enthousiasme, creativiteit en je geniale ideeën voor de meest uiteenlopende onderzoeken, waar de vrouw altijd een hoofdrol in speelt. Je hebt me geleerd trots(er) te zijn op mezelf en om mijn energie te steken in relaties en projecten waar ik zelf energie van krijg.

Beste **Charlotte, dr. Onland-Moret**, wat een fijn team had ik met jou en Hester in het UMC Utrecht. Jij hielp me om te structureren als ik echt niet meer wist waar ik heen moest met ~~mezelf~~ een manuscript of onderzoek. Op zo'n moment maakte jij altijd op korte termijn tijd voor overleg en hielp je bij structureren, prioriteiten stellen en een planning maken. Mede daardoor staat hier nu een prachtig proefschrift. Dankjewel!

Beste **Yolande, dr. Appelman**, je was een belangrijke steunpilaar voor mij in het VUmc/Amsterdam UMC. Het spijt me als ik je soms 'vergat' op de hoogte te houden, maar ik heb je kritische blik en feedback altijd heel erg gewaardeerd. Dit hielp om de manuscripten naar een hoger niveau te tillen. Ik bewonder je werk-ethos, maar ook de tijd die je neemt voor ontspanning (ook al zou voor veel mensen 568 km fietsen in Noorwegen geen ontspanning zijn).

Beste **Leo, prof. dr. Hofstra**, op de ESC van 2018 leerde ik je kennen als een bevlogen cardioloog. Ondanks je drukke tijden als cardioloog bij CCN, was het altijd prettig om tijdens de lunch even bij te praten en je op de hoogte te brengen van alles waar ik mee bezig was. Bedankt voor de mogelijkheid om te promoveren. **Yolanda**, heel hartelijk dank voor alle hulp bij het vinden van ruimte in Leo's agenda voor mij.

Graag wil ik de leden van de promotiecommissie bedanken, die de tijd genomen hebben om mijn proefschrift te beoordelen; **prof. dr. Piek**, **prof. dr. Verheij**, **prof. dr. ir. Isgum** en **dr. Cramer**. **Dr. Verjans**, **Johan**, bedankt dat je mij 5 jaar geleden introduceerde in de wondere wereld van het hart en de potentie van kunstmatige intelligentie. Het cirkeltje is rond nu jij ook mijn proefschrift beoordeeld hebt. Ook **prof. Bots**, **Michiel**, bedankt voor de beoordeling van mijn proefschrift en de leuke tijd in Tallinn. Ik bewonder hoe jij het hoogleraarschap invult!

Ernest Hemingway zei ooit 'The first draft of anything is sh*t'. Zonder **co-auteurs** waren ook de tweede, derde, vierde en vijfde draft zo gebleven. Ik wil hen graag bedanken voor hun bijdrage aan de verschillende artikelen in dit proefschrift. Een aantal co-auteurs wil ik in het bijzonder danken. **Dr. Aernout Somsen** en **dr. Igor Tulevski** van Cardiologie Centra Nederland, bedankt voor jullie feedback vanuit de klinische praktijk. **Rutger**, dankjewel! Het was heel prettig dat ik altijd even snel met je kon schakelen (in ons onderzoek met heel veel resultaten, maar zonder duidelijke boodschap). Ik ben heel benieuwd hoe jouw carrière-pad gaat verlopen, als MD met een schat aan programmeerkennis. **René**, bedankt voor de samenwerking en sparren over toekomstperspectief van TG'ers.
Beste **Birgitta**, **prof. Velthuis**, bedankt voor alle fijne mentorgesprekken die we gehad hebben. Het was fijn om zonder barrières te praten met iemand met een schat aan werk- en levenservaring over persoonlijke keuzes, werk-privé balans en gendergelijkheid.
Veel dank aan het CVON-AI team vanuit Groningen/Utrecht. **Jan-Walter**, wat was het fijn om zo nu en dan even met jou te sparren over onze PhDs, de academische wereld en het leven. Ik kijk uit naar jouw proefschrift, nog even! Beste **Pim**, **prof. dr. van der Harst**, bedankt voor je vertrouwen en je kritische blik op de onderzoeksresultaten van CVON-AI. **Prof. dr. van Rossum**, **Bert**, hartelijk dank dat ik in de VU kon promoveren. **Nicoline**, bedankt voor alle organisatie omtrent CVON-AI. **Gaby** en **Sandra**, bedankt voor alle organisatie vanuit het VUmc.
Studenten begeleiden vond ik één van de leukste componenten van het werk als onderzoeker. Ik heb veel geleerd van hen over goede begeleiding en heldere communicatie. **Hedy, Suzanne en Tamana**, bedankt voor de berg werk die jullie verzet hebben in jullie relatief korte stages. **Maxime**, ik ben trots op het paper dat we samen hebben kunnen schrijven op basis van jouw gedane (programmeer)werk. **Niels**, ik heb heel fijn met je samengewerkt naar een mooi paper. Ik zie je proefschrift over 4 jaar graag op de deurmat!
Science Lovers-collega's van de groep van Hester, bedankt voor alle weekstarts, taart en gezelligheid! **Ingrid**, jij houdt het lab draaiende! Bedankt voor alle gezelligheid en hulp bij het regelen van alle mogelijke zaken. **Gideon**, bedankt (voor de ontruiming van de Toren)! **Robin**, carrièretijger! Bedankt voor je gezelligheid! Ik weet nog een liedje; 't Is een... **Daniek**, **Mark**, **Tim**, **Elise**, **Ernest, Eliza** and **Michele,** bedankt voor alle gezelligheid, weekstarts, brainstormdagen en journal clubs. And good luck with your careers!
Anne-Mar, nu ben jij de senior PhD-onderzoeker! Nog een paar maandjes en dan kunnen we ook jouw proefschrift bewonderen. Bedankt voor alle fijne koffiemomentjes, meer digitaal (maar hee, dan kunnen we wel mooi de was opvouwen) dan live (hoe vaak die filterkoffie wel niet mislukt is…), maar altijd gezellig. Succes met de afronding van HelpFulUP, je PhD en je moestuin! **Diantha**, een mede-TG'er in de groep van Hester! Ik vind het heel knap hoe je middenin coronatijd begon met je PhD en het opzetten van een studie. Heel veel succes met de Epi-master en je promotieonderzoek. Ik kijk uit naar

jullie proefschriften en verdediging!

En dan de collega's van de 0900-Troostpot Hotline. **Jonne**, topper! Al voordat ik in het UMCU begon met werken, maakte jij het ernaar dat ik me thuis voelde. Recht voor zijn raap, met een hart van goud, het is echt prettig om met jou samen te werken. **Sophie**, ik viel op een gespreid bedje met een CCN dataset die al voor een heel groot deel opgeschoond was door jou (en waar al een jaar van je PhD-tijd in was gaan zitten!). Ik ben heel benieuwd waar de jaren jou gaan brengen, maar ik weet zeker dat het goed komt. Vergeet je niet voornamelijk werk te doen dat je leuk vindt? **Floor**, wederhelft in het dreamteam, perfect storm, partner-in-crime. Ik ben heel blij dat we elkaar gevonden hebben vanuit onze eigen expertise. Het resultaat: een paar prachtige papers en veel gezelligheid! We vullen elkaar goed aan en konden zo binnen no-time mooi werk neerzetten. Ik had me in jullie geen fijnere collega's kunnen wensen. Ook al scheiden onze wegen op professioneel vlak, ik hoop dat we onze koffies/thee/lattes/lunches/sushis/GiTo's erin blijven houden.

Na 1,5 jaar gewerkt te hebben op locatie in het ziekenhuis kwamen we helaas allemaal thuis te zitten. Maar in die 1,5 jaar heb ik heel veel plezier gehad op de Toren, de Tower of Power; **Michael**, **Malin**, **Saskia**, **Sander**, **Bram**, **Mark**! **PhD-collega's van de VU**, ondanks dat ik voor de coronatijd maar 1 keer per week in de VU was en daarna eigenlijk bijna nooit meer, voelde ik me altijd welkom. Bedankt voor de leuke uitjes, lunchpauzes, kerstdiners en koppen koffie! Ook wil ik graag alle leden van de **Asselbergs-group** en van het **Laboratory of Experimental Cardiology** bedanken voor de brainstorms, feedback-sessies en interessante presentaties.

Zoals de meeste mensen mij kennen, ben ik niet iemand die van stilzitten houd. Ook in mijn werk was ik continu op zoek naar nieuwe uitdagingen naast het reguliere PhD-onderzoek. Daarbij wil ik dan ook mijn collega's van **YoungSIT** bedanken voor onze inspirerende sessies op het gebied van Early Career Researchers, Erkennen&Waarderen en Open Science. Ook de mede-organisatoren van de **Circulatory Health Coffees** hebben mijn PhD-tijd tot een leukere tijd gemaakt.

Vrienden mogen ook niet ontbreken in dit dankwoord! Gezelligheid, een luisterend oor, samen sporten, wielrennen, hardlopen, koffiedrinken, een avondje eten. Zij zorgden ervoor dat ik kon ontspannen en met beide benen op de grond bleef staan. Mijn jaarclub Extase, zonnetjes van het mooiste en zonnigste studentenhuis in Enschede, mijn prachtige dispuut Pimpelle, sportmaatjes bij Hellas triathlon en van de Lombootycamp (weer of geen weer, wij staan er om 6.45), Utrecht me Homies, ofwel 'Enschede-kliek in Utrecht (en omstreken), oud-collega's van Q, de HIC tweeduidend én veertien, en vrienden van Skeuvel! Bedankt dat jullie er zijn.

In het bijzonder wil ik een aantal mensen bedanken: **Marleen**, **Alice**, **Rhodé**, **Marthe** en **Michelle**. Van samen zuipen in 't Gat tijdens onze allereerste Kick-In tot een kop thee

op de bank en de eerste baby's kruipend om ons heen. Ik kijk altijd weer uit naar onze etentjes! **Lieve zonnetjes van Noorderzon**, huisgenootjes van het eerste uur! Bedankt voor gezellige etentjes, koffie's en fietstochtjes. Mijn boekenclubvriendinnen: **Laura** en **Iris**. Ik geniet elke keer weer van onze maandelijkse boekbesprekingsavondje. **Sofieke**, wat is het heerlijk om met jou te sparren over de academie, toekomstplannen en veranderingen, of het nou bij Gys was of om kwart voor 8 's ochtends tijdens een rondje hardlopen. **Kicky**, zo gezellig dat je dichtbij woont en altijd in bent voor koffie om even bij te praten. Ik kijk uit naar jouw proefschrift en verdediging! Het NK Escaperoom-team, **Douwe**, **Adrienne**, **Marijn** en **Astrid**, online escaperoomen was een prachtig tijdverdrijf tijdens de corona. Ik kijk uit naar echte escaperooms! **Sietske** en **Tjalling**, ondanks dat jullie zo ver weg wonen, waardeer ik de moeite die jullie doen om langs te komen en op de hoogte te blijven! Het is altijd fijn en gezellig om bij jullie over de vloer te komen in Friesland. **Kevin**, ik ben blij dat ik je eindelijk ook een hardloper mag gaan noemen. Op tempo loop je me er inmiddels ruim uit, maar qua afstand heb je nog even te gaan. Binnenkort weer samen fietsen?

Lieve familie van Schless, ruim 2 jaar geleden kwam ik voor het eerst bij jullie over de vloer; wat een warm bad. Ik voelde me meteen thuis bij jullie in Roermond, maar ik zal mijn Limburgse vlaai met vorkje blijven eten. **Paul Sr**, **Marnel**, **Willemijn**, **Eduard**, **Steven** en **Eveline**. Ik hoop dat ik nog vele (sport)avonturen met jullie mag beleven. **Kiki** en **Lottie**, wat is het leuk om jullie van dichtbij te zien (op)groeien.

En dan mijn eigen thuisfront, dat giet fansels yn it Frysk. **Douwe**, broerke, it is altyd leuk mei dy! Bierkes, spultsjes, wille, kuierje en lekker ite. **Lotte** leart dy it Bourgondyske libben wol. Jimme binne altiten wolkom, sels yn Malaga! **Tynke**, myn paranimf, myn Yin fan'e Yang, myn sinne neist de moanne, myn Mont Blanc, myn dûnspartner. "The world was moving, she was right there with it and she was…" Do silst in prachtige takomst temjitte gean, mar ik wit ek hoe dreech it is om keuzes te meitsjen. It komt goed! **Jules**, ook jij zorgt voor de bourgondische gezelligheid bij onze familie. Bedankt (ook voor het feit dat je Tynke een dak boven haar hoofd geeft ;)). En dan (myn net sa) lytse broerke **Gerben**. ik ha dy meimakke fan lyts jonkje dat it leafst tomkjend tsjin de muziekbox harke nei Asterix en Obelix, oan't de yntelliginte jongeman dy'tst no bist. Ik bin benijd wêr de takomst dy bringt! **Christien,** tanke foar alles! Ik wit datsto net neamd wurde wolst, dus ik hâld it sa koart mooglik.

Heit en **Mem**, tanke foar al jimme leafde en stipe. Ik kin mij gjin bettere âlden winskje. Jimme hawwe ús altyd stipe, sûnder oardiel oer ús keuzes. It is altiten fijn en gesellich om thús te kommen en it is net foar neat dat Paul en ik graach tichtby Zeist wenjen bliuwe wolle. Ik hoopje, dat jimme stadichoan ek grutsk binne op jimme sels, hoe't jimme ús alle fjouwer nei ús plakje op dizze wrâld brocht hawwe.

En als laatste, lieve **Paul**, je wilde niet dat ik je zou bedanken, maar ik ben net zo eigen-

wijs als jij... Ik ben blij dat jij in mijn leven bent gekomen, je betekent de wereld voor mij en houdt mij op de been als ik het soms even niet zie zitten. Het leven met jou is fantastisch. Ik hoop dat hier nog heel veel jaren met campertripjes, reizen, klimsessies, hardloopavonturen, fietsritjes, klushuizen en zo nu en dan een avondje op de bank bij mogen komen.

CURRICULUM VITAE

Klaske Rynke Siegersma was born on July 18, 1991, in Utrecht, Netherlands, as the first child of four. She spent her childhood in Zeist, where she went to secondary school (Christelijk Lyceum Zeist). In 2009, she started the academic bachelor program Technical Medicine at the University of Twente. This educational program provided Klaske with a solid background in medical technology, medicine, patient care and science. In 2013, Klaske specialized in medical imaging and interventions during the 3-year master program of Technical Medicine, which she exchanged for a master in Health Sciences in 2014. From this program, she graduated cum laude with her thesis entitled 'Assessment of the added value and uncertainty of implementation of FDG-PET/CT and MRI-DWI in monitoring treatment response of stage-III non-small cell lung cancer'. In 2015 she proceeded with her clinical internships required to finish her degree in Technical Medicine. She conducted these clinical internships at the department of trauma surgery (Isala, Zwolle), department of orthopedics (UMC, Utrecht), Institute of Image Guided Surgery (IHU/IRCAD, Strasbourg) and department of cardiology (UMC, Utrecht). In her final internship, she discovered her love for the human heart, coding and the endless applications of artificial intelligence in our temporary healthcare. Klaske continued her graduation research in Technical Medicine at the department of Radiology at the UMC Utrecht. Therefore, she spent four months as an intern at the Computational Imaging and Bioinformatics Laboratory and Harvard Medical School. For this internship, she obtained a Dekker-scholarship from the Dutch Heart Foundation. In 2018, Klaske graduated cum laude for her masters degree in Technical Medicine with her thesis entitled 'Feasibility of radiomics in an MRI-dataset of patients with aortic stenosis to improve diagnostics, therapy and prognosis'.

In the same year, Klaske started her PhD research on the use of regular care data and development of early risk prediction tools in cardiovascular disease. This thesis is the result of a 3-year project. She performed her research at the department of cardiology at the Amsterdam UMC, location VUmc under supervision of prof. dr. Leonard Hofstra and dr. Yolande Appelman and in close collaboration with the laboratory of Experimental Cardiology at the UMC Utrecht, supervised by prof. dr. Hester den Ruijter en dr. Charlotte Onland-Moret.

Beside work, Klaske enjoys running, cycling and triathlon and loves the mountains. She finished multiple (mountain) marathons. When not outside, she likes to read books, cook and spend time with friends and family. After completing her PhD, Klaske aims to continue her career in big data analysis to improve healthcare and processes in healthcare.