Chapter 1
General introduction
& outline of the thesis
Population-based Screening

Developments within medicine over the last decades, such as new biomarkers and improved imaging methods, give the opportunity for early detection and treatment of diseases or conditions before they become symptomatic. This has led to population-based screening programs to be implemented for a variety of diseases in many high-income countries. In the Netherlands, eight different population-based screening programs are already implemented: Prenatal screening for infectious disease, Down, Edward’s or Patau’s syndromes, a structural ultrasound examination for foetuses, heel prick and hearing tests for neonates, and screening for colon, cervical and breast cancer\[1, 2\]. The number of deaths caused by cervical cancer was reduced by 24% in the Netherlands in 2006 when compared to the years 1986-1988, before the start of cervical cancer screening\[3\]. Similarly, it was estimated that the introduction of colon cancer screening in the Netherlands will have reduced mortality with 33% in 2030 and will prevent the occurrence of colon cancer with 20%\[4\].

The evidence for the necessity of screening for cardiovascular disease (CVD) is growing\[5, 6\]. A pilot-population screening for lung cancer and CVD has just been advised to be conducted by the Health Council of the Netherlands\[7\]. The results of the ROBINSCA trial show that determining coronary artery calcium (CAC) score could be used for population-based screening if it proves to be cost-effective\[8\]. Similarly, for lung cancer screening, the results of the US national lung cancer screening trial (NLST)\[9\] and NELSON\[10\] trial show reductions in mortality with early detection on CT and treatment, resolving points 1 to 7 of the Wilson and Jungner criteria for disease screening\[11\] (Table 1).

Screening increases the already substantial burden on radiology technicians and radiologists and takes finite human resources away from primary care. Not only because of the extra CT scans that need to be made and evaluated, but also because the evaluation of such scans involves the segmentation of structures, which is time-consuming. Therefore, another way of improving cost-effectiveness is to reduce the amount time necessary for evaluation of the screening test and improving its performance.

Table 1: Wilson and Jungner Criteria for Disease Screening\[11\].

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The condition of screening should be an important health problem</td>
</tr>
<tr>
<td>2.</td>
<td>There should be treatment for patients diagnosed with the disease</td>
</tr>
<tr>
<td>3.</td>
<td>Facilities to diagnose and treat the disease should be available</td>
</tr>
<tr>
<td>4.</td>
<td>There should be a recognizable latent or early symptomatic stage</td>
</tr>
<tr>
<td>5.</td>
<td>A suitable test or examination should be available</td>
</tr>
<tr>
<td>6.</td>
<td>The test should be acceptable to the population</td>
</tr>
<tr>
<td>7.</td>
<td>The natural history of the condition should be adequately understood</td>
</tr>
<tr>
<td>8.</td>
<td>There should be agreement in the policy of whom to treat as patients</td>
</tr>
<tr>
<td>9.</td>
<td>The cost of screening, diagnosis, and treatment should be economically balanced within the total cost of health care spending</td>
</tr>
<tr>
<td>10.</td>
<td>Screening should be a continuing endeavour to allow for refinement in screening methods, outcomes, and process improvement</td>
</tr>
</tbody>
</table>
A way of reducing the workload for technicians and radiologists could be to automate some of the processes within the screening workflow with artificial intelligence (AI) tools. The field of AI in medical imaging has rapidly advanced the last decade due to new breakthroughs in GPU accelerated deep learning (DL) and in efficiency of convolutional neural networks (CNN). The range of problems that deep learning can be applied to is numerous and varied and are often categorized into classification, detection, segmentation and regression tasks\(^\text{(12)}\). Examples include automatic contouring of organs\(^\text{(13)}\), improving image quality\(^\text{(14)}\), or outcome prediction\(^\text{(15)}\). For screening, AI algorithms could assist physicians by finding biomarkers for diseases that humans cannot detect, creating new tests and evaluations, furthering points 4 and 5 of the Wilson and Jungner criteria (Table 1). Similarly, AI has been helping with discovering and developing drugs\(^\text{(16)}\), and promoting point 5 by creating new treatments for diseases. The advancement of technology and increased interest in AI in the last decade resulted in better accuracies of deep learning methods\(^\text{(17, 18)}\). The primary aim of this thesis is to contribute to the automation of the test stage of CVD screening with CAC scoring on low-dose CT by using deep learning methods.

**Use case: Cardiovascular disease screening**

In 2017, 19.9 million new cases of CVD were diagnosed in Europe, 3.6 million of which caused by ischaemic heart disease (IHD)\(^\text{(19)}\). CVD cost the EU 210 billion euros in 2015. These costs can be divided in 53% direct healthcare, 21% informal care costs, 15% productivity loss due to mortality and 11% productivity loss due to morbidity. IHD accounts for 59 billion euros of the expenditure on CVD. This burden is estimated to increase in the following years\(^\text{(20)}\).

Atherosclerosis is often the underlying cause of IHD. Remodelling of the coronary arteries can lead to deposits of fatty plaques building up in the vessel wall. Although the first stages only involve wall thickening, narrowing of the vessel can happen at a later stage. This narrowing of oxygen bearing blood vessels can cause ischaemia of the heart muscle, which can develop, if too much tissue is damaged, into a myocardial infarction. The early stages of atherosclerosis can happen asymptptomatically, making early detection difficult without proper screening methods.

Therefore, to combat the rising impact of CVD, the European Society of Cardiology in 2016 recommend to systematically screen people at higher risk for CVD as estimated by the systematic coronary risk evaluation (SCORE)\(^\text{(5, 21)}\). However, the SCORE is developed for individual use, not for population wide screening. Another method for risk classification is scoring the coronary artery calcium by using the Agatston score\(^\text{(22)}\). The Agatston score has been shown to correlate with future cardiovascular events decades ago. Although developed in 1990 on Electron Beam Computed Tomography (EBCT)\(^\text{(23)}\), recent studies have shown that the Agatston score can be translated to dual source ECG-triggered CT\(^\text{(24, 25)}\). Monosource
CT systems are not reliable when calculating the CAC score and might cause stratification errors when used for AI applications. The Agatston or CAC score is calculated by identifying and segmenting CAC spots on dual source ECG-triggered CT and multiplying the size of the spots with a factor based on the highest density of each spot. In the ROBINSCA and ImaLife trials, focusing on CAC screening and on CAC, lung cancer and COPD screening, respectively, the CAC scoring is done semi-automatically by trained readers using dedicated software (CaScoring, Syngo.via, version VB30A, Siemens Healthineers). Although individual scoring does not take much time, a minute per patient without CAC and about 10 minutes per patient with CAC, the high volume of patients needing scoring makes the time required by technicians substantial.

The Multi-Ethnic Study of Atherosclerosis (MESA) shows that CAC scoring improves risk classification and, with these results, CAC scoring is projected to improve cost-effectiveness of CVD screening\[^{26}\]. The first results of the ROBINSCA trial have shown that the CAC score would reduce the number of people at risk, thereby needing less preventive treatment to be needed\[^{8, 27}\]. Further results of the ROBINSCA and ImaLife trials might give insight in whether screening with CAC is effective enough to be implemented. By applying deep learning to the test, we might be able to reduce the number of false positives and decrease the time necessary for evaluation of the test, thereby increasing cost-effectiveness even further.

**Evaluation and deployment of AI algorithms**

Due to the improved capabilities of AI tools and availability of both AI software libraries and large datasets, research on using AI tools for clinical tasks is being done in almost all specialisations and departments of academic hospitals all over the world. The number of papers published on deep learning methods such as CNNs for medical tasks have risen exponentially since 2015\[^{28}\]. However, the evaluation, deployment and monitoring of AI tools is still lacking governance. Similarly, the problems of biases, protection of privacy and trust of both public and medical professionals in AI tools are largely unaddressed\[^{29}\].

The European council and parliament approved new medical device regulations (MDR) in May 2017 to become effective starting 26 May 2021. One of the key changes is that medical devices, including software, need continuous post market evaluation. Regular updates on clinical evaluation, safety and performance are required. Another key change is that most software as a medical device (SaMD) has been given a class II medical device status instead of class I, therefore, no longer classifying SaMDs as low risk devices and now requiring a notified body assessment\[^{30}\]. AI tools will perform differently on populations everywhere it will be deployed, due to usually being trained on a certain population. In 2019, the guidelines on ethics for trustworthy AI were published by the European Commission. This document defines trustworthy AI as lawful, ethical and robust. According to these
guidelines, one of the basic principles of trustworthy AI should be non-discrimination. This means that AI algorithms should work equally well on different population groups. However, this is contradictory with another key point in these documents: Ensure data privacy. It is of importance to have a very broad dataset for evaluations of AI tools. In the proposal of the European Commission for a regulation laying down harmonised rules on artificial intelligence, it is stated that AI systems need to be validated and tested on datasets that have appropriate statistical properties regarding the cohorts they are intended to be used on[31]. However, international datasets are difficult to acquire due to privacy laws.

This problem is further enhanced by the ability of AI medical software to continue updating by learning from user cases, thereby changing by use. It is, therefore, critical to create guidelines for validation to increase transparency for the sake of implementation of AI tools. Next to this, evaluation of these SaMDs relies heavily on the use of quantitative metrics, such as sensitivity, specificity, odds ratio, positive and negative predictive values, and confidence intervals, each with small differences towards which type of error they are sensitive. Comparisons between such metrics are often difficult without having the classification or segmentation ground truth and it remains uncertain if these metrics correlate with clinical performance.

Outline of this thesis

The primary aim of this thesis is to contribute to the automation of the test stage of CVD screening with CAC scoring on low-dose CT by using deep learning methods. A secondary aim of this thesis is to investigate how to evaluate such deep learning tools so that it can be implemented responsibly.

The first two chapters describe the development process for AI software in general and more specifically for cardiac CT image analysis. Chapter 2 describes the types of applications and challenges in medical AI tool development. Different types of tasks an AI algorithm could assist with and an overview of the stages the development process of machine (ML) and deep learning software go through before implementation are characterised. To ensure the quality of an AI tool and prevent biases creeping in, it is important to have a well-constructed structure while designing, training and validating the tool. Chapter 3 goes further into detail on the types of uses for ML and DL algorithms within the field of cardiac CT. It provides an overview of the types of AI tasks with recent applications within cardiac CT and describes requirements for implementation of such applications. The problems with the current state of medical AI research are elaborated on and suggestions are given for improvement. In Chapter 4, the development of an automatic pipeline for exclusion of CAC in low dose CT is illustrated as the first use case in this thesis. The process of building a deep learning architecture and its training and internal and external validation are characterised. The
reasoning for selecting the process of CAC scoring for automatization is further elaborated on and a comparison is made to similar work. Moreover, the implications are described on implementing this pipeline in the clinical workflow. Chapter 5 describes the second use case, a deep learning algorithm to automatically contour the heart and substructures. Development of this pipeline is done to expand and improve the pipeline for CAC scoring described in chapter 4. Different deep learning methods for segmentation tasks are shown and tested for this particular topic.

To investigate its use in clinical practice, a qualitative evaluation is done on the heart contours as seen in Chapter 6. Here, the reasons why it is difficult to compare results of medical AI papers is discussed, how the field of medical AI research could evolve to improve this and whether overlap metrics are useful to describe how close an AI tool is to implementation in the clinic. Finally, Chapter 7 summarizes all the chapters, debates the broader perspective of this thesis and presents the implications for future research.
References

3. RIVM (2010) Effecten van vaccinatie en screening


30. Medical Device Coordination Group (2020) MDCG 2020-1 Clinical Evaluation of Medical Device Software
