# LUND UNIVERSITY

**Artificial intelligence in PET-CT. From Image Enhancement to Imaging Biomarkers.**

Ly, John

2020

*Document Version:*
Publisher's PDF, also known as Version of record

[Link to publication](#)

*Citation for published version (APA):*
Ly, J. (2020). *Artificial intelligence in PET-CT. From Image Enhancement to Imaging Biomarkers.* [Doctoral Thesis (compilation), Department of Translational Medicine]. Lund University, Faculty of Medicine.

*Total number of authors:*
1

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

# Artificial intelligence in PET-CT

## From Image Enhancement to Imaging Biomarkers

**JOHN LY**
**DEPARTMENT OF TRANSLATIONAL MEDICINE | LUND UNIVERSITY**
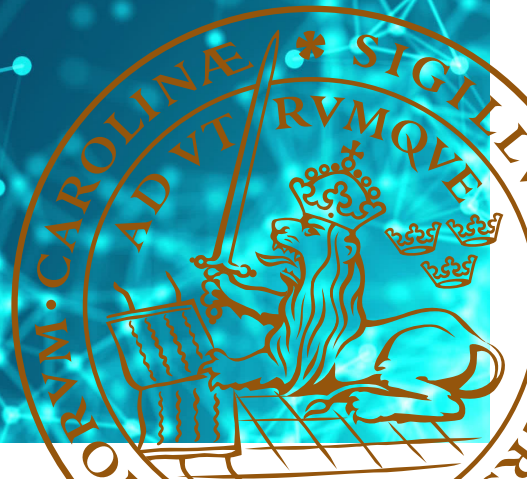
Artificial intelligence in PET-CT

# Artificial intelligence in PET-CT

## From Image Enhancement to Imaging Biomarkers

**John Ly**

LUND
UNIVERSITY

DOCTORAL DISSERTATION
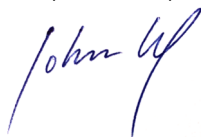
*Faculty opponent*
**Professor Heber MacMahon**
University of Chicago, United States

| **Organization**<br>LUND UNIVERSITY | **Document name**<br>DOCTORAL DISSERTATION |
|---|---|
| Faculty of Medicine<br>Department of Translational Medicine<br>Clinical Physiology and Nuclear Medicine, Malmö | **Date of issue** |
| Author: John Ly | Sponsoring organization |

| **Title and subtitle:** Artificial intelligence in PET-CT - From Image Enhancement to Imaging Biomarkers |
|---|

**Abstract**

The rapid growth of positron emission tomography with computed tomography (PET-CT) usage and the vast amount of quantifiable data from its examinations have led to an increased demand in software which can quickly detect, analyse and provide clinical decision support on findings. The aim of this thesis is to investigate if the use of various custom-made artificial intelligence (AI) algorithms can enhance the diagnostic power of PET-CT in patients with lung cancer and lymphoma.

Paper I investigated whether a basic AI algorithm had good agreement with and was faster than human measurements of maximum standard uptake values ($SUV_{max}$) in serial PET-CT examinations. In this retrospective study, 26 lesions were included from lung cancer and lymphoma patients for manual and semi-automatic $SUV_{max}$ measurements. There was good agreement between the methods and the semi-automatic method was up to five times faster than the manual method.

Paper II investigated whether choice of PET reconstruction algorithm affects the classification of tumour response, Deauville score (DS). AI was used to calculate $SUV_{max}$ in reference organs. Three reconstruction methods were chosen: Q.Clear (QC, General Electric's proprietary algorithm), and the European Association of Nuclear Medicine Research Ltd. (EARL) recommendations (pre-2019 update) at its upper and lower limits for the resolution recovery coefficient. In this prospective study, 52 patients with non-Hodgkin and Hodgkin lymphoma were included. There were significant differences in DS between the QC and both EARL reconstructions but not between the upper and lower limits of the EARL reconstructions.

Paper III developed and explored an AI algorithm's sensitivity in lung tumour detection and its ability to calculate total lesion glycolysis (TLG). In this retrospective study, 112 patients were recruited. The sensitivity for the AI algorithm was 91%. The positive predictive value and negative predictive value was 88% and 100%, respectively on a patient level. The TLG agreement was good and higher in smaller lesions.

Paper IV used an AI algorithm to enhance PET-CT images. 72 patients were recruited for training and 25 for testing of the algorithm. The AI algorithm was able to reduce noise and increase contrast compared with standard images whilst keeping $SUV_{max/peak}$ stability. In conclusion, the AI algorithms used in this thesis were able to function as a complement to and increase the quality of PET-CT examinations.

| **Key words**<br>Artificial intelligence, PET-CT, segmentation, reconstruction algorithm, lymphoma, lung cancer |
|---|
| Classification system and/or index terms (if any) |

| Supplementary bibliographical information | **Language English** |
|---|---|
| **ISSN** and key title    1652-8220 | **ISBN** 978-91-7619-987-9 |

| Recipient's notes | **Number of pages** 60 | Price |
|---|---|---|
| | Security classification | |

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature    *John Ly*                            Date 2020-10-16

# Artificial intelligence in PET-CT

## From Image Enhancement to Imaging Biomarkers

John Ly

LUND UNIVERSITY

# Table of Contents

# List of papers

This thesis is based upon the following papers which will be referred to by their Roman numerals in the text.

I. **Ly J**, Garpered S, Höglund P, Jönsson E, Valind S, Edenbrandt L and Wollmer P : Semi-automatic analysis of standard uptake values in serial PET/CT studies in patients with lung cancer and lymphoma BMC Medical Imaging 2012; 12:6

II. **Ly J**, Minarik D, Edenbrandt L, Wollmer P and Trägårdh E : The use of a proposed updated EARL harmonization of $^{18}$F-FDG PET-CT in patients with lymphoma yields significant differences in Deauville score compared with current EARL recommendations EJNMMI Research 2019; 9:65

III. Borrelli P*, **Ly J***, Kaboteh R, Ulén J, Enqvist O, Trägårdh E, Edenbrandt L : AI-Based Detection of Lung Lesions in [$^{18}$F]FDG PET-CT from Lung Cancer Patients (Submitted)

IV. **Ly J**, Minarik D, Jögi J, Wollmer P, Trägårdh E : Post-reconstruction enhancement of [$^{18}$F]FDG PET images with a convolutional neural network (Submitted)

Published papers are reprints with kind permission from the publishers.

*Shared first authorship.

# Abbreviations

| | |
|---|---|
| AI | Artificial intelligence |
| BSREM | Block sequential regularized expectation maximization |
| DL | Deep learning |
| DS | Deauville score |
| EANM | European Association of Nuclear Medicine |
| EARL | European Association of Nuclear Medicine Research Ltd |
| [$^{18}$F]FDG | Fluorodeoxyglucose |
| ML | Machine learning |
| OSEM | Ordered subset expectation maximization |
| PET-CT | Positron emission tomography with computed tomography |
| PMT | Photomultiplier tube |
| QC | Q.Clear © |
| SiPM | Silicon photomultiplier |
| SUV | Standard uptake value |
| TLG | Total lesion glycolysis |

# Summary

The rapid growth of positron emission tomography with computed tomography (PET-CT) usage and the vast amount of quantifiable data from its examinations have led to an increased demand in software which can quickly detect, analyse and provide clinical decision support on findings. The aim of this thesis is to investigate if the use of various custom-made artificial intelligence (AI) algorithms can enhance the diagnostic power of PET-CT in patients with lung cancer and lymphoma.

Paper I investigated whether a basic AI algorithm had good agreement with and was faster than human measurements of maximum standard uptake values ($SUV_{max}$) in serial PET-CT examinations. In this retrospective study, 26 lesions were included from lung cancer and lymphoma patients for manual and semi-automatic $SUV_{max}$ measurements. There was good agreement between the methods and the semi-automatic method was up to five times faster than the manual method.

Paper II investigated whether choice of PET reconstruction algorithm affects the classification of tumour response, Deauville score (DS). AI was used to segment reference organs in which $SUV_{max}$ was calculated. Three reconstruction methods were chosen: Q.Clear (QC, General Electric's proprietary algorithm), and the European Association of Nuclear Medicine (EANM) Research Ltd. (EARL) recommendations (pre-2019 update) at its upper and lower limits for the resolution recovery coefficient. In this prospective study, 52 patients with non-Hodgkin and Hodgkin lymphoma were included. There were significant differences in DS between the QC and both EARL reconstructions but not between the upper and lower limits of the EARL reconstructions.

Paper III developed and explored an AI algorithm's sensitivity in lung tumour detection and its ability to calculate total lesion glycolysis (TLG). In this retrospective study, 112 patients were recruited. The sensitivity for the AI algorithm was 91%. The positive predictive value and negative predictive value was 88% and 100%, respectively on a patient level. The TLG agreement was good and higher in smaller lesions.

Paper IV used an AI algorithm to enhance PET-CT images. 72 patients were recruited for training and 25 for testing of the algorithm. The AI algorithm was able to reduce noise and increase contrast compared with standard images whilst keeping $SUV_{max/peak}$ stability.

In conclusion, the AI algorithms used in this thesis were able to function as a complement to and increase the quality of PET-CT examinations.

# Populärvetenskaplig sammanfattning
# (Summary in Swedish)

PET-CT är en undersökningsmetod som kartlägger anatomisk (CT) och funktionell information (PET) samtidigt genom konventionella röntgenstrålar och radioaktiva läkemedel som injiceras i kroppen i samband med undersökningen. Användningsområdena utgörs framför allt av diagnostik och uppföljning av tumörer men kan även användas för t.ex. utredning av inflammatoriska/infektiösa tillstånd och demens.

Den snabba utvecklingen av PET-CT och den stora mängd data som går att utvinna från dess undersökningar har lett till en efterfrågan på programvaror som snabbt kan upptäcka, analysera och ge kliniska beslutsstöd av fynden. Syftet med avhandlingsarbetet är att undersöka om användandet av olika egenutvecklade algoritmer som baseras på artificiell intelligens (AI), kan förbättra diagnostiken i PET-CT-undersökningar hos patienter med lungcancer och lymfom.

Studie I undersökte om en enkel AI algoritm hade god samstämdhet och om det fanns skillnader i mättiden jämfört med mänskliga mätningar av maximum standard uptake value ($SUV_{max}$) i seriella PET-CT-undersökningar. I denna retrospektiva studie inkluderades tio lungcancer- och sex lymfompatienter som utfört två till fyra PET-CT undersökningar. Tjugosex tumörer inkluderades för $SUV_{max}$ mätning med manuell och halvautomatisk metod. Tre granskare mätte $SUV_{max}$ med båda metoder och tidsåtgången för respektive metod. Samstämdheten mellan metoderna och mellan granskarna var goda och den halvautomatiska metoden var upp till fem gånger snabbare än den manuella metoden.

Studie II undersökte om olika PET-CT rekonstruktionsalgoritmer påverkar indelningen av behandlingssvar hos lymfompatienter på den så kallade Deauville-skalan. De tre specifika algoritmerna som undersöktes var: Q.Clear (QC, General Electrics proprietära algoritm) och rekommenderade inställningar vid övre respektive nedre gränsvärde enligt European Association of Nuclear Medicine Research Ltd. (EARL). I denna prospektiva studie inkluderades 52 patienter, varav 18 kvinnor och 34 män, med lymfom. Signifikant skillnad sågs mellan QC och de rekommenderade inställningarna enligt EARL, men inte mellan de övre och nedre EARL-rekonstruktionerna. Fem patienter som granskats med QC klassificerades som icke-responders men vid användning av övre och nedre EARL rekonstruktionerna klassificerades dessa som responders vilket hade lett till annorlunda klinisk bedömning och handläggning.

I studie III utvecklades en AI-algoritm i PET-CT-undersökningar för att upptäcka lungtumörer och automatiskt beräkna total lesion glycolysis (TLG), en kvantitativ mätenhet för tumörbörda, hos patienten. I denna retrospektiva studie inkluderades 112 patienter, 59 kvinnor och 53 män, som delades in i en träningsgrupp (n = 66),

valideringsgrupp (n = 23) och testgrupp (n = 23). AI-algoritmen hade god förmåga att hitta individuella lungtumörer och missade inga patienter med lungcancer. Detta möjliggör prioritering av vilka undersökningar som bör besvaras först.

Studie IV berörde bildförbättring av PET-CT-undersökningar med hjälp av AI, där olika inställningar i rekonstruktionsalgoritmen och insamlingstider jämfördes med en AI-förbättrad inställning. 97 patienter deltog i studien, varav 72 användes för att träna AI-nätverket och 25 patienter utvärderades med olika bildrekonstruktioner med och utan AI-förbättring. Det fanns ingen signifikant skillnad i $SUV_{max}$ eller $SUV_{peak}$ i patologiska fynd mellan standard och AI-förbättrad rekonstruktion med standardinsamlingstid. Brus- och kontrastnivån bedömdes vara bättre i de AI-förbättrade bilderna. Resultaten kan omsättas i förbättrad bildkvalitet, minskad insamlingstid och/eller minskad stråldos för patienten.

Den övergripande slutsatsen i denna avhandling är att AI-algoritmerna i studierna kan användas som ett komplement och öka kvaliteten på PET-CT-undersökningar.

# Introduction

Artificial intelligence (AI) has made incredible progress in the last decade due to continuous advancements in computational power and deep learning algorithms such as convolutional neural networks (CNN) which excels in image-recognition tasks [1, 2]. The technology can alleviate the need for imaging specialists to perform repetitive and time-consuming tasks, which is an inefficient use of intellect that could instead be directed towards concluding findings. Although research advancements have been made by utilising AI in the medical field, e.g. in image enhancement [3-5], segmentation [6-9], classification and prognostication [10-14], the actual implementation of these in clinical practise in a fully automated fashion pose several challenges, but are nevertheless believed to take place in the future [15].

[$^{18}$F]FDG PET has several advantages over CT alone, it provides quantifiable metabolic data which is useful in appreciating disease activity, oncological staging and follow-up examinations [16]. The added benefit of combining PET with CT is the provision of high-resolution anatomic information and attenuation correction factors. The drawback of PET is the added ionising activity for the patient, the limited availability and additional cost of radiopharmaceutical tracers, but advancements have been made in deriving standard-dose CT [17] and PET images [18] from low-dose data using CNNs.

This thesis utilises AI in all of its constituting papers and depicts the departure from basic algorithms to modern advanced machine- and deep learning methods. It aims to explore the feasibility of using AI, from enhancing imaging studies to calculation of total lesion glycolysis, an emerging imaging biomarker, for research and clinical application.

# Background

## Artificial intelligence

Artificial intelligence is a broad term used to demonstrate computer logic and advanced functions. It ranges from simple logic made of if-statements in a spreadsheet in Excel to machine learning (ML), which online stores use to recommend your next purchase and deep learning (DL) that enables autonomous cars to recognise objects on the road. The distinction between ML and DL lies in the details. While both methods concern the learning ability of AI, ML algorithms are dependent on manually provided features whereas DL can learn features independently but requires much larger datasets and is considered as a subset of ML.

*Applications in medical imaging*

The most basic form of AI implementation in medical imaging could be the clinical decision support tools ACR Select [19] and iGuide [20]. These tools are tailored by experts based on the appropriateness criteria developed by the American College of Radiology. Logic programming is used to suggest radiologic examinations depending on the patient information that is provided by the physician. Although not touted as AI tools in their marketing, perhaps rightfully so, they would fall under the broader definition of AI for the layman, whereas experts primarily associate AI with machine and deep learning [21]. Other early and basic forms of AI in medical imaging are segmentation and co-registration algorithms, which have been refined lately with deep learning algorithms.

Advancements in AI was possible in the 2000s due to increasing computational power and access to larger datasets. A breakthrough was made in 2013 when a few graphics processing units was able to replace thousands of computer processing units to train CNNs [22]. Early ML algorithms were referred to as computer-aided detection systems and were typically used to solve pattern-recognition problems such as the detection of nodules or calcifications. These have evolved into DL based algorithms which are better at generalising disease representations.

In this thesis, Paper I utilises basic AI whereas Paper II-IV use convolutional neural networks which is a DL algorithm that is specialised in image recognition and image enhancement tasks. There are many other deep learning algorithms which are used in other areas, e.g. face recognition, machine translation and cyber security.

*The structure of a CNN*

A convolutional neural network imitates how the brain (neurons) is structured (network) [1, 2]. Convolution is a mathematical operation that is performed on data through filters. CNN architectures vary depending on what task is at hand. For an image classification CNN (Figure 1), an image serves as the input to the CNN where several convolutions in numerous layers takes place and is passed on to a pooling layer where the data are subsampled – this saves computational power and avoids overfitting, a modelling error which reduces the CNNs generalisability. The steps taken up until now is called feature extraction, e.g. edges and corners. The features are sent to the fully connected layers (neurons that are interlinked) where classification is performed through the activation of the "neurons". The combination of neurons that are activated decides what the final result will be in the output layer, which typically is a set of classifiers e.g. the numbers 0-9, animals, fruits, type of organs, lesion or not lesion.
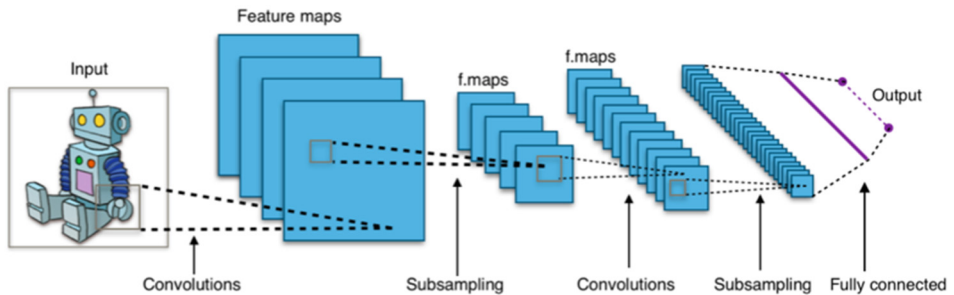


**Figure 1. Schematic of a convolutional neural network used for image classification.**
CNN layers between the input and output layers are often referred to as "hidden layers" because their functions are not observable. ("Typical CNN architecture" by Aphex34 is licensed under CC BY-SA 4.0)

*Perceptions of AI in medical imaging*

It is believed by experts that AI will continuously improve medical imaging in many facets such as radiation dose, image reconstruction and quality, detection, classification, monitoring, reporting, workflows and even interventional radiology [15, 21, 23, 24]. Since ML applications are more transparent than their DL counterparts, the way to regulatory approval is more rigorous for the latter. A common misconception is that AI will replace imaging specialists. This is likely founded in low understanding of the current capabilities of AI and the legal obstacles that need to be overcome due to the nature of DL algorithms. Ongoing education within imaging departments regarding the advancements in AI is important to prepare for the inevitable transformation AI will bring.

# PET-CT imaging

*Imaging technique*

PET-CT is a nuclear medicine technique which combines PET with the x-ray CT, which provides functional and anatomical information, respectively. The resulting images from the modalities are superimposed to provide superior diagnostic accuracy, mostly in oncological applications [25].

For PET, the patient is injected with a radiopharmaceutical which is a drug containing a radioactive isotope. These are used to diagnose different types of conditions and diseases in the body and are also called tracers. The most commonly used tracer is fluorodeoxyglucose ([$^{18}$F]FDG), which is a glucose analogue coupled with the radioisotope fluorine-18 ([$^{18}$F]) and basically shares the same kinetics as glucose. Where there is elevated glucose metabolism, such as in inflammatory/infectious sites and cancer cells in the body, accumulation of the tracer in these cells will happen more than in other normal cells which enables the differentiation during interpretation of the PET-studies.

The tracer undergoes continuous positron decay while it is distributed in the body. The scanning takes place after a delay which depends on which tracer is used. For [$^{18}$F]FDG, an interval of 60 min after injection is recommended [26] but can vary for other tracers due to differing pharmacokinetics.

When the positron interacts with an electron, an annihilation event occurs, from which a pair of high-energy photons travel in opposite directions – the photons that reach the detectors of the PET scanner within the coincidence time window will become the basis of the PET image (Figure 2). Two other types of coincidence events can occur. Random coincidence occurs when two annihilation events are detected as one due to a lost or undetected photon from each annihilation. Scattered coincidence happens due to Compton scattering of the photons in the patient body where the photons change direction. Both of these latter types of coincidence events lead to incorrect spatial detection of the original annihilation event and needs to be corrected.

The detector consists of scintillation crystals coupled with either photomultiplier tubes (PMT) or newer generation silicon photomultipliers (SiPM) (Figure 3). The latter are smaller, require less voltage and electromagnetically insensitive – making them compatible with PET-magnetic resonance systems. In comparison studies, PET-CT systems with SiPM were more sensitive and had higher image quality than those without SiPM [27-29].

When a photon enters a crystal and interacts, the previous high-energy photon is converted into lower energy photons that are measured by the PMT/SiPM. In the next step, these measurements are organised in a histogram matrix from which image reconstruction can occur.
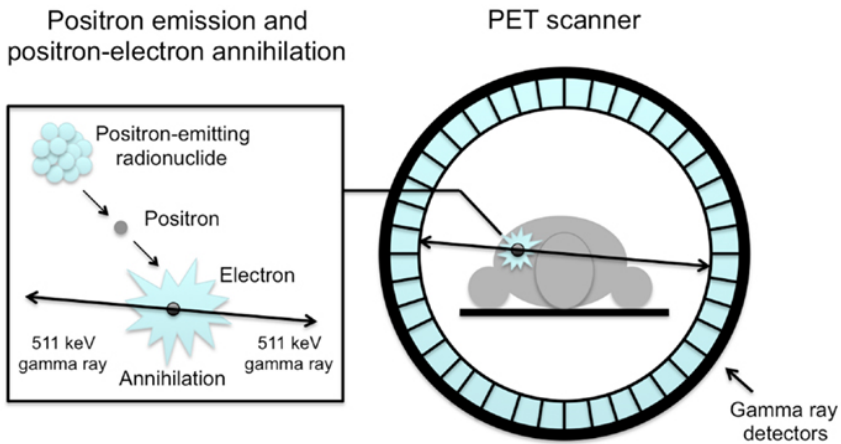
**Figure 2. Detection of an annihilation event in a PET scanner.**
The placements of the detectors in rings with the scintillation crystals pointed towards the paitent enables the estimation of where the annihilation event occurred.
("Positron emission and positron-electron annihilation" by Adriaan Lammertsma is licenced under CC BY 3.0)



**Figure 3. Images of photomultipliers used in PET.**
Left: A photomultiplier tube. Right: Silicon photomultipliers with different array sizes.
(The images have been published with courtesy of Hamamatsu Photonics K.K.)

*Image reconstruction methods*

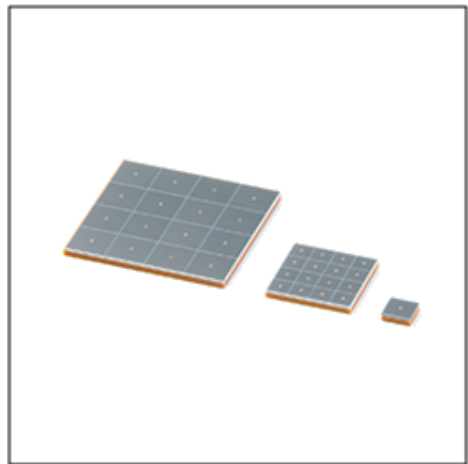Filtered back projection was for a long time the most common method for image reconstruction which provides fast reconstruction, but does not take noise and physical effects into account [30]. It has since been replaced by iterative reconstruction methods, which allow modelling of the mentioned weaknesses into the reconstruction and are today the most widely used technique for reconstructing PET-CT images. In this method, a measured image is compared to the projected image which is then updated and used iteratively in several loops until the image best represents reality. The number of iterations needed to render objects in the image visible depends on the object size with smaller objects requiring more iterations. This means that the resolution in the image increases with more iterations. The noise in the image, however, also increases with the number of iterations. The noise in the image is usually regularized with a post reconstruction smoothing filter, often a Gaussian filter, which always reduce the resolution in the image. There is thus a trade-off between the resolution and noise when choosing the number iterations and post filter used in the reconstruction. Since several advanced calculations are performed in each iteration, this technique was not readily available in the past nor feasible in a clinical context due to its long processing time. One of the first popular iterative reconstruction methods was maximum likelihood expectation maximization, which was later replaced by ordered subset expectation maximum (OSEM) where the image is divided into subsets for calculations [31]. It was found that the image quality was comparable with the added bonus of far less computational effort [32].

Novel iterative reconstruction methods such as block sequential regularized expectation maximization (BSREM), which is used in half of the papers in this thesis, has developed the iterative reconstruction method further by taking neighbouring voxels into account. Earlier methods would treat each voxel independently from each other. Compared to traditional reconstruction algorithms, BSREM can maintain a low noise level as the number of iterations increases. The algorithm suppresses noise via a penalty factor β, in which higher values suppress noise more but also reduce resolution. BSREM has been shown to have better quantitative accuracy of SUV, particularly in small lesions [29, 33-36] compared to OSEM with post-filtering. Figure 4 demonstrates different iterative reconstruction algorithms.

The use of deep learning in conjunction with image reconstruction in clinical practise is limited, but there are studies which uses CNNs to simulate standard-dose CT [17] and PET images [18, 37] derived from low-dose data.
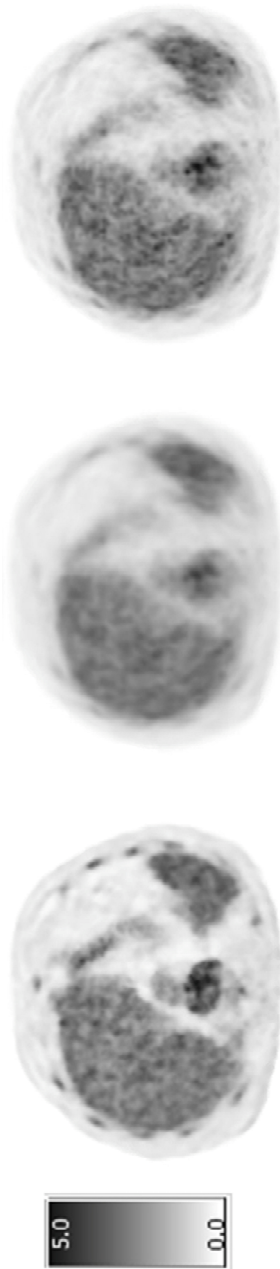
**Figure 4. Transaxial PET images of the liver with different iterative reconstruction algorithms.**
Left: BSREM algorithm. Middle and Right: OSEM algorithm with different iteration, subset and post filter settings. The middle image was reconstructed with fewer iterations and filtered with a wider post filter than the right image. (Screenshots by Elin Trägårdh.)

## Segmentation

Delineating an organ or hotspot, by separating it from the background on an image is a process called segmentation. This can be done manually in a computer software ranging from free hand tools to circular (two-dimensional) region of interest (ROI) or volumetric (three-dimensional) volume of interest (VOI). There are different automatic methods utilising AI to choose from, which have varying performance depending on the situation. The more basic methods include graph-cut [38] (Figure 5), thresholding, clustering, active contours and adaptive region growing. The advanced methods utilise CNNs such as DeepMedic and U-Net [39].
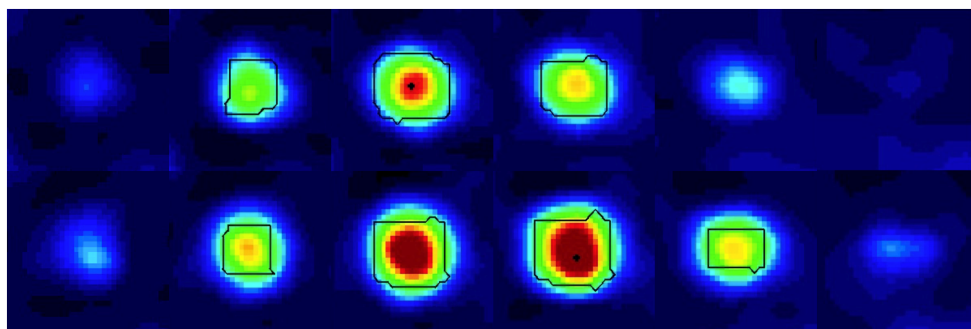


**Figure 5. Semi-automatic segmentation of hotspots in PET images.**
A hotspot is segmented in two serial PET-CT examinations (CT not shown) using the graph-cut algorithm. Top row is the first examination and bottom row is the follow-up examination. The black dot is a manual indication which helps the graph-cut algorithm in finding the borders (black polygonal circuit) of the hotspot and to find the corresponding hotspot in the second examination.

## Technical harmonisation

To ensure image comparability between machine vendors and encourage multicentre research, EARL has an accreditation programme with two specification standards for $[^{18}F]$FDG SUV recovery coefficients – the original "standard 1" which was published in 2010 [40] and updated recommendations in 2019 "standard 2" [41, 42], which takes recent advancements in PET hard- and software into consideration as these have an effect on SUV as previously described. The need for harmonisation in methodology for hotspot segmentation and texture analysis, also known as radiomics, is widely recognised and further efforts in these areas have been called upon [43-45].

# Lymphoma

Lymphocytes are cells of the immune system which are found most frequently in the lymph nodes, spleen, thymus and bone marrow. When lymphocytes are genetically changed and grow out of control, which is the definition of cancer, the condition is called lymphoma. There are two broad classifications of lymphoma. Hodgkin lymphoma has two peak incidences, young people in their 20s and people in their 70s. Non-Hodgkin lymphoma is the most prevalent and has an increasing incidence with age which peaks around 80 [46].

PET-CT is primarily used for staging and assessment of treatment response in lymphoma. Malignant lesions are typically characterised by focally elevated activity within enlarged lymph nodes but may in some cases have a diffuse pattern within organs or in the abdominal fat which is more difficult to detect and appreciate.

Treatment is usually but not limited to chemotherapy, among other treatments are radiation therapy, bone marrow transplantation and immunotherapy. The 5-year survival rate is good compared to other cancers, for Hodgkin lymphoma it is 87% and for non-Hodgkin lymphoma it is 73% regardless staging [47].

*Deauville score*

To assess interim or end of treatment response on PET-CT in lymphoma patients, a five-point scale was adopted in 2009 at the first international workshop on PET in lymphoma which was held in the French city Deauville [48]. Generally, DS 1-3 signifies metabolic response and DS 4-5 is considered as no metabolic response, see Table 1. The benefits of this scale, which was confirmed in large trials [49, 50], is the possibility to predict outcome in patients after two cycles of chemotherapy and use this knowledge to escalate or de-escalate treatment accordingly [51]. It is also useful in predicting 3-year progression-free survival in patients with diffuse large B-cell lymphoma [52] and follicular lymphoma [53]. Also, the relapse rate in natural killer T cell lymphomas after chemo- and chemoradiotherapy [54]. The scale can be used visually or quantitatively comparing the uptake in tumour, liver and mediastinal aorta.

**Table 1. Deauville five-point scale.**
A score from 1 to 5 is assigned according to tumour uptake in relation to reference organs.

| Score | Definition |
| --- | --- |
| 1 | No uptake |
| 2 | Uptake < mediastinum |
| 3 | Uptake > mediastinum but < liver |
| 4 | Moderately increased uptake compared to the liver |
| 5 | Markedly increased uptake compared to the liver and/or new lesions |
| X | New areas of uptake unlikely to be related to lymphoma |

# Lung cancer

The most preventable cancer disease and the leading cause of cancer deaths in the world is lung cancer [55]. Smoking cessation reduces the risk of getting lung cancer and premature death [56]. Broadly, lung cancers are divided into two types; small cell and non-small cell lung cancers, the latter being the most common. The tumour cells almost always arise from the bronchi or alveoli. If detected early, it often has an isolated round appearance, with or without additional characteristics such as spiculation, within the lung parenchyma or sometimes adjacent to the hila or mediastinum. To further characterise these lesions, [$^{18}$F]FDG PET-CT can be performed where elevated metabolic activity can be appreciated but some tumour subtypes require other radioactive tracers to be detected.

Staging of lung cancer is commonly done with PET-CT, which is especially beneficial to patients with non-small-cell lung cancer where more than half of the patients obtain a different staging [57] and one out of five patients avoid unnecessary surgery [58] compared to workups not using PET. TLG, calculated as $SUV_{mean}$ x metabolic tumour volume, is an emerging imaging biomarker that can predict outcome in patients with NSCLC [59, 60].

Treatment is usually but not limited to surgery, among other treatments are chemo-, radio-, immuno- and laser therapy. The 5-year survival rate is 21% [47].

# Rationale

The usage of PET-CT examinations has undergone rapid growth since the turn of the millennium, the first PET-CT system was used in 1984 at Gunma University in Japan [61]. The number of performed [$^{18}$F]PET-CT examinations per year has been growing 5-15% annually in Sweden for the last ten years [62]. Being a quantitative modality, physicians need to assess the activity, often comparing with reference organs in order to appreciate if the activity corresponds to a pathologic process. Sometimes the so-called hotspots of suspected disease have similar uptake compared to the reference organs which necessitates more accurate measurement and comparisons of SUVs, which is time consuming and require precise segmentation.

In lymphoma monitoring, Deauville score is used to determine the treatment response in tumours, the score is based on the comparisons of several SUVs. TLG is an emerging tumour marker which has shown to be useful in the prognostication and evaluation of treatment response in cancer patients [63]. To calculate TLG, precise tumour volume and SUV measurement is needed.

The repeated use of PET-CT examinations poses a risk for the patient in terms of radiation, this is particularly true for patients with malignant melanoma, colorectal and anal cancer, who usually undergo multiple scans for staging, treatment response and follow-up after end of treatment [64-68].

The use of computational power and AI could potentially alleviate these problems and is explored upon in this doctoral thesis with the included studies. The focus has been placed on investigating feasibility of the usage of custom-made AI methods in imaging situations, where physicians can use these tools to make their workflow more effective, obtaining precise measurements of SUV and TLG, enhancing image quality and provide benefits for the patient by reducing radiation.

# Objectives

The main objectives of this thesis were:

1) To demonstrate the time saving quality and reproducibility using a semi-automatic AI method for calculation of $SUV_{max}$ of abnormal lesions in serial PET-CT studies compared with manual method.

2) To investigate whether using a novel state-of-the-art SiPM-based PET-CT with QC reconstruction (which complies with the 2019 EARL recommendation) affects DS compared with reconstructions meeting the previous EARL harmonizing standard in patients with lymphoma.

3) To develop a completely automated AI method for the detection of lung cancer in [18F]FDG PET-CT images and automatically measure the TLG compared with manual measurements.

4) To investigate if image quality of BSREM reconstructed images obtained with short and standard acquisition times can be improved using a CNN trained on images acquired with a long acquisition time.

# Materials and methods

## Study populations

This thesis recruited patients from research databases established by Skåne University Hospital Malmö/Lund and Sahlgrenska University Hospital in Gothenburg, Sweden. The examinations in the databases consists of clinically indicated examinations from patients that have provided written consent to be included in the database.

In Paper I, patients with lung cancer or lymphoma who had undergone two serial [18F]FDG PET-CT studies between July 2008 and January 2010 at Skåne University Hospital in Malmö were included retrospectively. Inclusion criteria were defined for pathological lesions: 1. Sharp contrast to surrounding areas 2. No formation of a large conglomerate mass 3. Presence of lesion in both examinations. A total of 16 patients were included, 10 lung cancer and 6 lymphoma patients. Each patient had one to four pathological lesions, which in total was 26 lesions.

In Paper II, patients diagnosed with lymphoma who had undergone an [18F]FDG PET-CT study between November 2017 and March 2018 or August 2018 to October 2018 at Skåne University Hospital in Malmö or Lund were included prospectively. The split dates were due to a shortage of cases in the first period and so a decision was made to collect more. Furthermore, our research databases do not store examinations in list-mode by default, therefore we were unable to recover patients between the two time periods. Patients referred for baseline PET, interim PET, end of treatment PET and suspicion of recurrence were included. Patients were required to have SUV uptake in the lymphoma lesions on PET (DS 2-5) and a corresponding lesion on CT. A total of 52 patients were finally included in the study.

In Paper III, patients who underwent clinically indicated [18F]FDG PET-CT due to suspected lung cancer or for the management of known lung cancer between April 2008 and December 2010 at Sahlgrenska University Hospital in Gothenburg were included retrospectively. In the selection process, three patients were excluded because of centrally located tumour, likely sarcoid disease and mediastinal tumour. A total of 112 patients were included and was divided into a training group (59%; n=66), a validation group (20.5%, n=23) and a test group (20.5%; n=23).

In Paper IV, patients referred for clinical [18F]FDG PET-CT at Skåne University Hospital, Malmö or Lund, were included in the study. Seventy-two of the patients

were included prospectively (December 2019 to March 2020) and used for training the CNNs. A separate set of 25 patients was included retrospectively (April to June 2018) and were used to evaluate the algorithms.

# PET-CT

All patients underwent an intravenous injection of 4 MBq/kg body weight of [$^{18}$F]FDG after at least 4 hours fasting and at a glucose level $\leq$ 10 mM. Imaging was performed 60 minutes after administration from the mid-thigh/inguinal region to the base of the skull. The patients were scanned with 5-10 bed positions, depending on the length of the patient.

In Paper I, one Philips Gemini TF was used. Acquisition time per bed position was 2 min with blob-ordered-subset time-of-flight, a row-action maximum-likelihood algorithm, as PET reconstruction method.

In Paper II, three Discovery MI (GE, Healthcare, Milwaukee, WI, USA) were used. Acquisition time per bed position was 1.5 min with QC (ß500) as PET reconstruction method.

In Paper III, one Siemens Biograph 64 Truepoint was used. Acquisition time per bed position was 1.5 min with OSEM as PET reconstruction method.

In Paper IV, four Discovery MI (GE, Healthcare, Milwaukee, WI, USA) were used. For the CNN training group, one of the bed positions had an acquisition time of 6 min and the rest were 1.5 min each (the 6 min acquisition was used for training). The images of the patients in the test group were acquired with a time per bed position of 4 min for all bed positions and stored in list-mode. BSREM was used as PET reconstruction method.

# Basic AI

In Paper I, basic AI-tools were used. To co-register CT images, a rigid registration algorithm was used, the CT images were reduced in size, blurred, thresholded and then overlaid to compare pixels with the formula:

$$\sum_{(i,j)} erf(i,j) = |IM_1(i,j) - IM_2(i,j)| \qquad (1)$$

One of the images is repositioned and the formula is applied again to calculate the pixel error. The process is reiterated with less size reduction and blurring each time.

The reader marked an arbitrary pixel of a pre-determined pathological lesion in one of the two PET studies. After this manual step, the program automatically detects the corresponding lesion in the second PET study. The reader's mark in the first PET study has a corresponding location in the first CT study, provided that PET and CT images were correctly aligned from the PET-CT camera. The corresponding location in the second CT study was defined using the matching by the software and the corresponding location in the second PET study was defined.

The volumetric segmentation of the pathological lesions was made in the PET images. The reader's mark in the first PET study and the corresponding location in the second PET study were used as seed points for the segmentation using the graph-cut algorithm. $SUV_{max}$ in the entire lesion was then calculated automatically and the segmentations were presented to the reader. Semi-automatic analysis was done on a standard laptop computer.

# Advanced AI

The Research Consortium for Medical Image Analysis (RECOMIA) has an online-platform which is used to facilitate to collaboration between medical and AI-researchers with the use of AI-tools to aid in scientific experiments [69]. Among the platform's multiple functions, annotation of organs and pathology, quantification and deep learning-based tools were relied upon in Paper II and III.

For Paper II, the tool was used to semi-automatically segment the liver and mediastinal blood pool and manually segment lymphoma lesions in order to calculate Deauville scores. For Paper III, manual segmentation of lung lesions was used as gold standard and compared with the AI's segmentation and TLG.

*AI models*

Two CNNs were used in Paper III, one which segments organs and one that detects lesions. The organ segmentation CNN (Figure 6) has an architecture inspired by the popular U-Net [39]. In the pre-processing step of the CT images, Hounsfield values are clamped to $[-800, 800]$ and divided by 800, resulting in an input with values in the range $[-1, 1]$. To train the CNN, 339 CT examinations with approximately 13.000 manual organ segmentations were divided into a training (80%) and validation (20%) set. The final post-processing step for all organ labels consists of extracting the largest connected component and filling holes in that component.
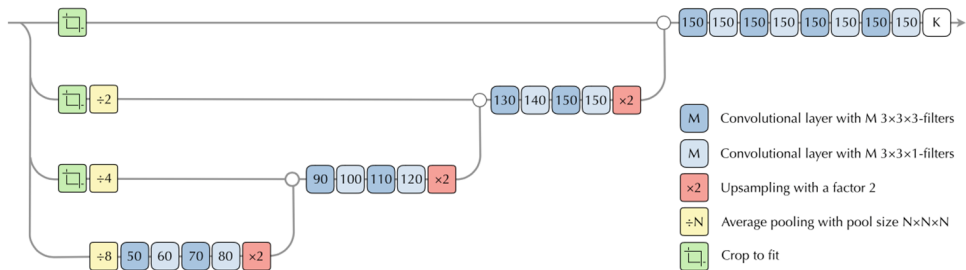
**Figure 6. The convolutional neural network structure used in RECOMIA.**
Two different filter sizes are used to compensate for anisotropic voxel sizes and producing an approximately cubic field of view. (Image by Olof Enqvist, reproduced with permission.)

The Detection CNN was trained the same way as the previously described CNN. Since exact delineation of lesions is virtually impossible, mimicking the exact boundaries of the annotations is not relevant. Thus, any voxels within 10 mm from the annotated lesions are marked as "don't-care". This means that when computing the loss function, there is no loss for these voxels regardless of the output label. For the remaining voxels the standard negative log-likelihood loss was used. Naturally, this leads to a slight over-segmentation of the lesions, but as detection was the main goal, this was considered acceptable.

In Paper IV, a denoising CNN suitable of image enhancement [70] with 10 convolution layers with a 256 x 256 x 5 matrix input was used. Each convolution layer consists of 68 3 x 3 x 3 filters except the first and last layer which consist of one 3 x 3 x 3 filter. A linear rectifier was used after each convolution layer except the last one. A mean squared error loss and a stochastic gradient descent optimizer was used.

List-mode data from the bed position with a 6 min scan time were extracted from each of the 72 patients in the training group. From the list mode data, 1 image set for 6 min and 4 image sets for 1 (first, second, third and fourth minute) and 1.5 min (6 min divided in 4 intervals) was reconstructed with the BSREM algorithm, respectively. For the 1 and 1.5 min images, a ß of 500 was used, and a ß of 200 was used for the 6 min image. Two sets of training pairs were composed, pairs with 1 and 6 min images and pairs with 1.5 and 6 min images. 10 pairs of training images were extracted from the 50 centremost slices from all patients – these were further divided into 10 subsets, each comprising a 3D volume sized 256 x 256 x 5. Furthermore, each reconstruction (1 and 1.5 min) were randomly resized, sheared and flipped in two dimensions, resulting in 5 additional samples. The total training pairs for each of the two training sets were (subsets x patients x reconstructions x samples) 10 x 72 x 4 x 6 = 17.280. Two networks were trained: one for images acquired with 1 min and the other for 1.5 min/bed position.

# Statistical analyses

All statistical analyses were performed by the principal author aided by the co-authors using IBM SPSS Statistics (Armonk, NY, USA). Bland-Altman plots were produced in Microsoft Excel (Redmond, WA, USA). In all papers, continuous variables and their distributions are presented as means with standard deviations (SD) and categorical variables are characterized by percentages or numbers.

In Paper I, intraclass correlation was used to calculate interobserver reproducibility. Bland-Altman analysis was conducted to assess the level of $SUV_{max}$ agreement between segmentation methods.

In Paper II, Friedman's test was used to compare DS between the three reconstruction methods for $SUV_{max}$ and $SUV_{peak}$, respectively. Significant p-value was set at $p<0.05$. Post-hoc Wilcoxon ranked-mean test was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.0167$.

In Paper III, Bland-Altman analysis was conducted to assess the level of TLG agreement between segmentation methods. Correlation between manual and AI-based TLG was assessed using Pearson correlation coefficient. The two analyses were repeated after removing an outlier.

In Paper IV, Friedman's test was used to compare quantitative data for $SUV_{max}$, $SUV_{peak}$ and COV measurements in all 5 reconstructions. Significant p-value was set at $p<0.05$. Post-hoc analysis with Wilcoxon signed-rank test was conducted with a Bonferroni correction applied, resulting in a significance level set at $p<0.005$. Since there were 5 reconstructions, each measurement group resulted in 10 comparisons and thus significant p-value was calculated as $0.05/10 = 0.005$.

Kruskal-Wallis test was applied to each investigator's results for the investigation of potential difference between the groups. Post-hoc testing with Mann-Whitney U tests were performed on all pairwise groups for each reader which resulted in 10 comparisons per reader. Thus, when Bonferroni correction was applied significant p-value was calculated as less than $0.05/10 = 0.005$.

# Ethics

All the studies presented in this thesis were carried out in accordance with the Declaration of Helsinki. Paper I, II and IV were approved by the Regional Ethical Review Boards at Lund University. Paper III was approved by the Regional Ethical Review Boards at Gothenburg University. All patients provided written informed consent.

# Results

*Main findings of Paper I*

Bland-Altman's 95% limits of agreement were 0.46 to -1.88; 0.77 to -1.92 and 0.23 to -1.31 for each reader (Figures 7-9). Manual and semi-automatic method agreed in all cases whether $SUV_{max}$ had increased or decreased between serial studies. The average time to measure $SUV_{max}$ changes in two serial PET-CT examinations was four to five times longer for the manual method compared with the semi-automatic method for all readers (reader 1: 53.7 vs. 10.5 s; reader 2: 27.3 vs. 6.9 s; reader 3: 47.5 vs. 9.5 s; $p < 0.001$ for all).
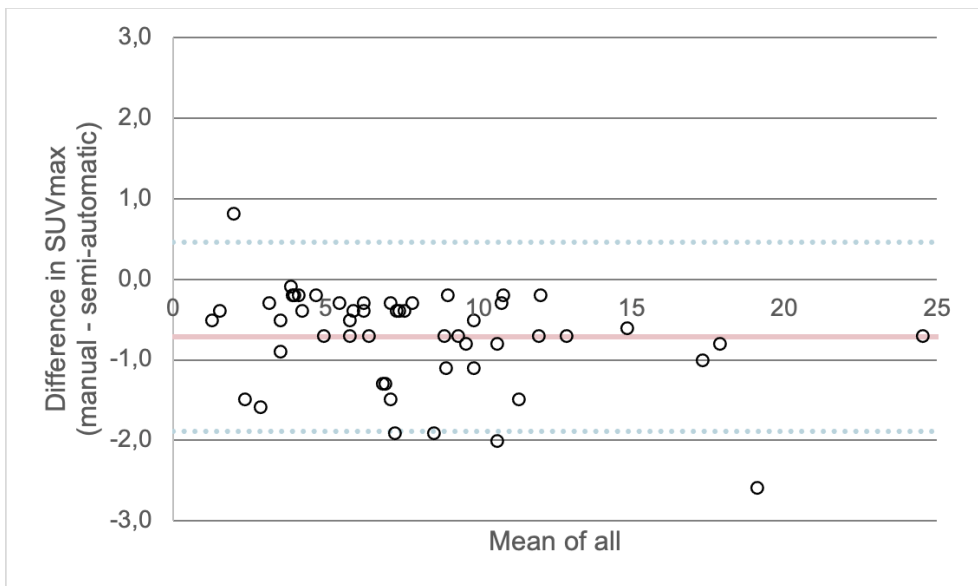


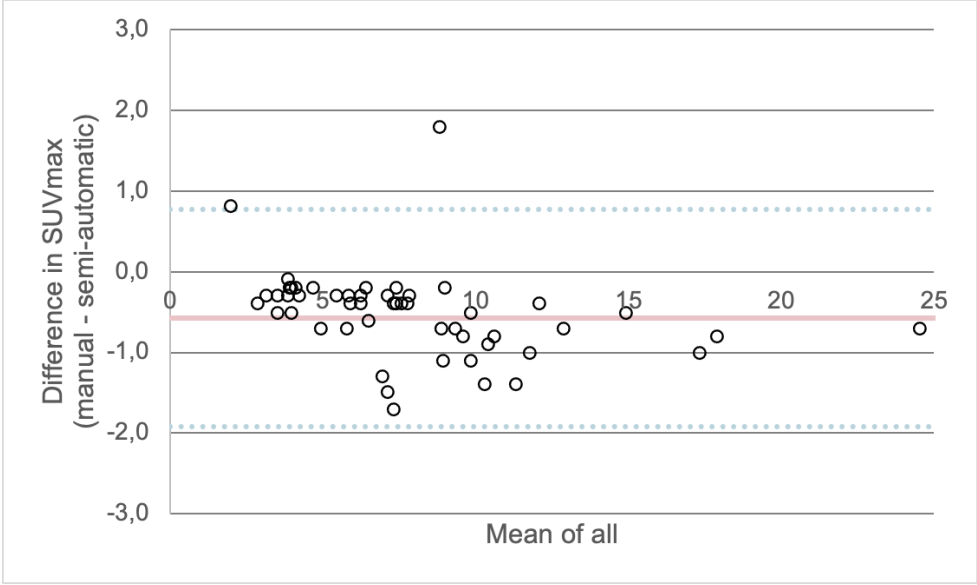**Figure 7. Bland-Altman plot for reader 1.**
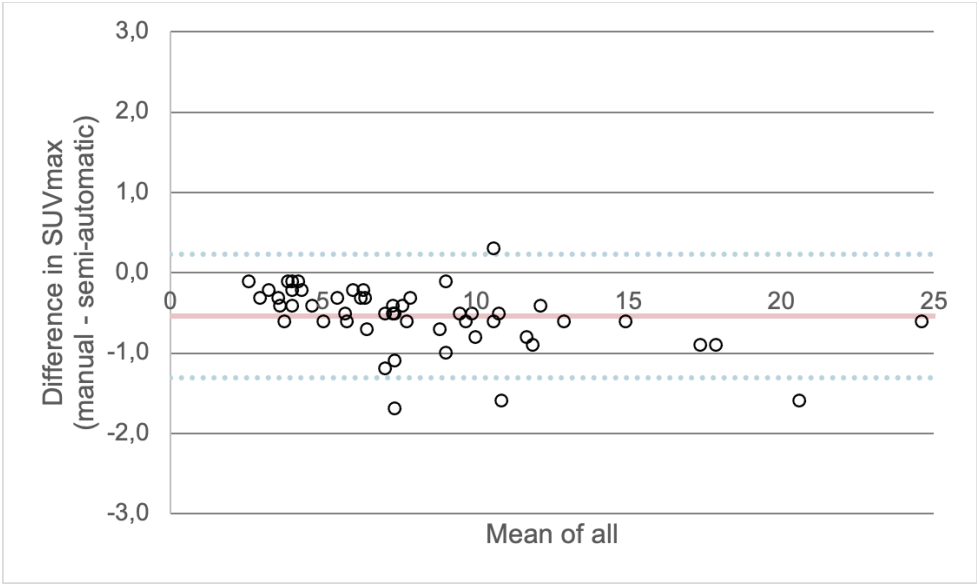
**Figure 8. Bland-Altman plot for reader 2.**



**Figure 9. Bland-Altman plot for reader 3.**

*Main findings of Paper II*

There was a significant difference in DS between the QC algorithm and EARL$_{lower}$/EARL$_{upper}$ (p < 0.0001 for both) but not between EARL$_{lower}$ and EARL$_{upper}$ (p = 0.102) when SUV$_{max}$ was used. For SUV$_{peak}$, there was a significant difference between QC and EARL$_{lower}$ (p = 0.001), but not for QC vs EARL$_{upper}$ (p = 0.071) or EARL$_{lower}$ vs EARL$_{upper}$ (p = 0.102). Five non-responders (DS 4–5) for QC were classified as responders (DS 1–3) when EARL$_{lower}$/EARL$_{upper}$ was used, both when SUV$_{max}$ and SUV$_{peak}$ were investigated. A graphical overview is seen in Figure 10.
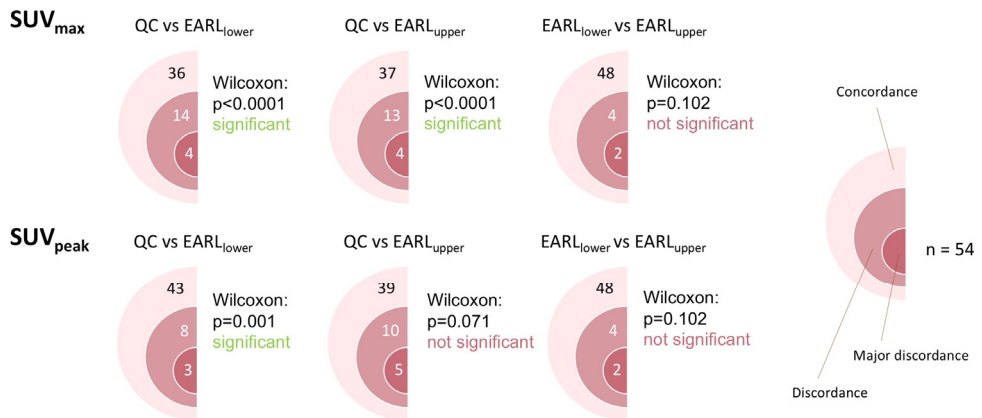


**Figure 10. Diagrams of pairwise comparisons between reconstruction algorithms.**
Numbers in coloured semi-circles represent n cases that had concordance, discordance and major discordance, respectively, when comparing reconstruction algorithms pairwise. For SUV$_{max}$, post-hoc analysis with Wilcoxon signed-rank test with Bonferroni correction, QC compared to EARL$_{lower}$ and EARL$_{upper}$ respectively yielded significant p value but not for EARL$_{lower}$ compared to EARL$_{upper}$. For SUV$_{peak}$, only QC compared to EARL$_{lower}$ yielded significant p value.

*Main findings of Paper III*

The AI-tool's performance in detecting lesions had a sensitivity of 91%. One small lesion was missed in two patients where both had a larger lesion which was correctly detected. The positive and negative predictive values were 88% and 100%, respectively, on a patient basis. The correlation between manual and AI TLG measurements was strong ($R^2 = 0.74$). Bias was 42 g and 95% limits of agreement ranged from -736 to 819 g. Agreement was particularly high in smaller lesions. Figure 11 shows an example of a tumour that was difficult for the AI to segment, probably due to the complex nature of the tumour, including areas of necrosis.
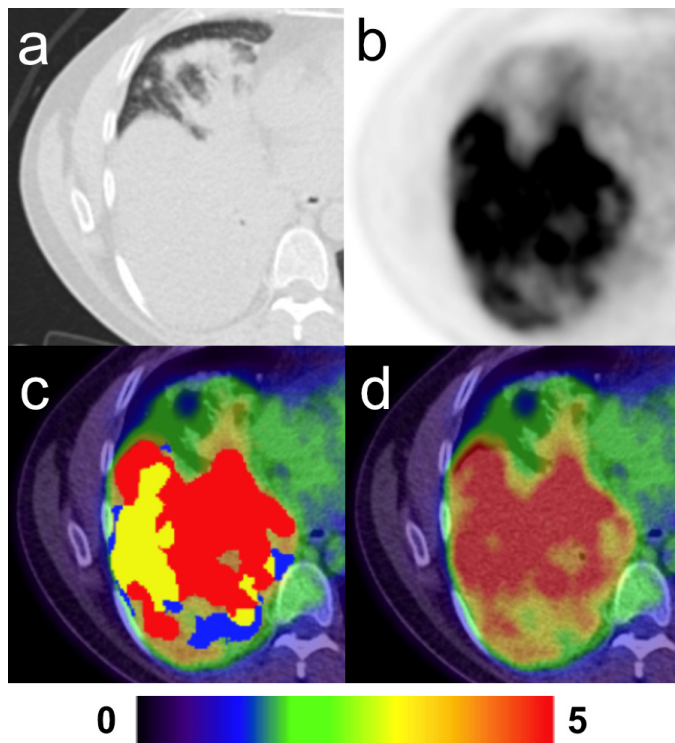


**Figure 11. Large lesion with complex necrosis in the right lung.**
a) axial CT with lung window. b) axial PET. c) fused axial PET-CT with overlaying segmentations; manual only (red), AI-only (blue) and manual + AI (yellow). d) fused axial PET-CT.

*Main findings of Paper IV*

There was no significant difference in hotspot $SUV_{max/peak}$ between the standard 1.5 min and 1.5 min CNN images. Coefficient of variation (COV), i.e. the noise level, was lower in the CNN enhanced images compared with standard 1 min and 1.5 min images. Physicians ranked the 1.5 min CNN and the 4 min images highest regarding image quality (noise and contrast) and the standard 1 min images lowest. Examples of the image sets are seen in Figure 12.
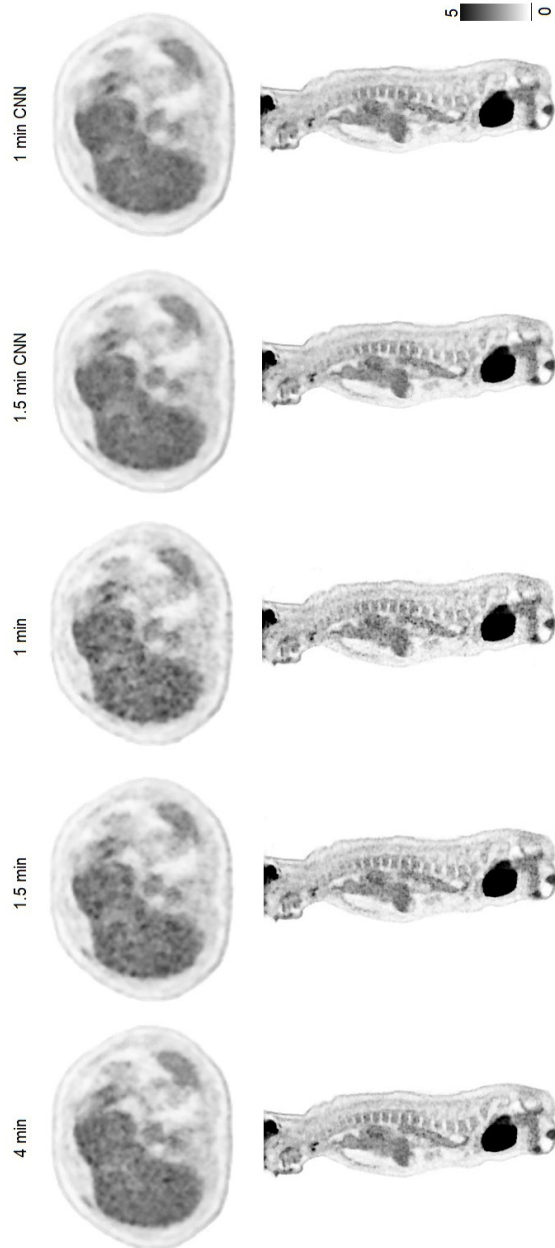
**Figure 12. Transaxial PET images of the upper abdomen and midline sagittal images of the image sets in the test group.**
Noise in the liver increases as scan time per bed position is decreased, both corresponding CNN series have markedly less noise in the liver. (Screenshots by Elin Trägårdh.)

# General discussion

There is a demand for objective and standardised methods when it comes to quantifying imaging studies [43]. The literature shows that when radiologists evaluate findings differently it inevitably leads to divergent patient management [71]. The evolvement of new guidelines by international imaging societies may sometimes not be fully adopted by the institutes but instead undergo local adjustments for different reasons and in the end, there may still be problems with individual adoption and compliance [72]. The use of AI-tools could potentially ameliorate these problems.

Since the publication of the first paper of this thesis (8 years ago), computational power has continued to increase exponentially in accordance with Moore's law [73]. Quantification of data in PET exams has shown to aid in the characterization of morphology, staging, treatment evaluation and prognosis [44, 74]. Some of these data cannot be extracted by normal reading of the exam by a physician. AI can therefore not only aid the physician with normal tasks but also provide individual patient insight that is normally not available.

## Clinical implications

Manual quantification of SUV metrics in PET-CT examinations is a repetitive and time-consuming task. SUV in lesions is often compared with reference organs which have variable values intrinsically and in between examinations. Paper I showed that AI was both faster and retained interobserver reproducibility and showed good agreement for its $SUV_{max}$ measurements compared with manual method. This finding would enable secure and efficient measurements if introduced clinically, which was not available at that time. Since the study was published, commercial software suites have incorporated this type of functionality.

With the introduction of newer PET scanners and reconstruction methods, special care must be considered with SUV measurements. In Paper II, the comparison between old and new EARL recommendations [75] led to different DS scoring which in turn could impact clinical management. Clinics that adopt new technologies must be aware that the literature surrounding DS is based on old technologies and thus the newer reconstruction methods are not validated with the

use of DS. Subsequent studies have demonstrated that the newer EARL recommendation results in increased SUVs up to 30% [76], which suggests that it could lead to changed DS classification compared with older EARL recommendations. To date, there is no clear evidence whether it is best to use DS obtained by newer hard- and software (such as the new EARL recommendations) or with older ones (such as the former EARL recommendations). Further clinical studies on patient outcome are needed to answer this question.

PET-CT examinations are read by physicians that are specialised in radiology and / or nuclear medicine, sometimes two specialists read the PET and CT separately – this may lead to backlogs of unread examinations which causes a delay in patient management. In Paper III, AI was successful in obtaining 100% negative predictive value for lung cancer on a patient basis. This is beneficial in situations when prioritisation of examinations is needed, which saves time and improves patient care.

The radiation exposure from an $[^{18}F]$FDG PET-CT examination is higher than what a CT examination would be. The limited spatial resolution and inherent noisiness of PET images are two factors that can contribute to difficult assessments. In Paper IV it was discovered that AI-enhanced PET images compared with standard images (ß500, 1.5 min scan time per bed position) outperformed in qualitative metrics (less noise and increased contrast) whilst retaining quantitative ($SUV_{max/peak}$) metrics. This opens up the possibility of providing better images to physicians or reducing acquisition time/administered activity.

# Methodological discussion and limitations

SUV from manual and AI measurements can sometimes have inherent inequalities. In Paper I, manual measurements were done on a Philips workstation with accompanying Philips software suite where images are interpolated, whereas the AI-tool was created in MATLAB without interpolation of images which most likely resulted in a systematic error. This is unavoidable as it is not possible to obtain the details of the post-processing Philips use in their software.

Measuring how long it takes to extract $SUV_{max}$ from pre-defined lesions is only a fraction of what it actually takes a physician to read a PET-CT exam but demonstrates a common repetitive task for nuclear medicine physicians. A digital stopwatch was controlled by each reader to record the time which introduces the possibility of small variations between the start of the clock and measurement and between obtaining the measurement and stopping the watch. The sample size was likely enough to smooth out these variances. Also, the expected difference (more than double) allows for the use of a digital stopwatch.

Selection bias of lesions that were small to medium in size with clear boundaries likely contributed to the almost-perfect reproducibility for both manual and AI measurements. Other studies at the time [77, 78] had similar inclusion criteria due to the known fact how segmentation tools were functioning at the time.

Paper II is unique when it comes to calculating Deauville score. AI was used to semi-automatically segment the reference organs completely (excluding lesions), instead of placing an arbitrary ROI/VOI in reference organs or evaluated visually which is usually done clinically. The latter methods may be difficult to assess if tumour-to-organ ratio is near 1:1. Complete segmentation of the organs increase the probability of calculating the true $SUV_{max}$ or $SUV_{peak}$.

Similar studies [79, 80] have not found a difference in DS classification when comparing OSEM with newer technologies such as PSF and reconstruction algorithms e.g. BSREM/QC. The problem is that within older and newer reconstruction algorithms it is possible to set parameters that influence the noise level and thus also SUV. It is known that BSREM and modern hardware improvements such as time-of-flight and point-spread function increase SUV [27], so it was natural to suspect that it would also affect DS classification. There are limitations in our study, firstly the sample size was small and monocentric. Secondly, the indication for the PET-CT exam did not matter for inclusion in order to increase sample size even though Deauville score is mostly used in interim PET and end of treatment PET exams.

It would be of interest to compare the upper and lower limits of the new EARL recommendations, but for our PET-CT system, longer acquisition times are necessary to reach the new upper limit, which was not feasible for the current study.

In Paper III, TLG agreement between AI and manual calculation was analysed with Bland-Altman plots in order to visualise if agreement was dependent on TLG levels. Disparities may arise if the AI has included FDG-avid structures adjacent to the lesion or in tumours with complex necrosis pattern. We chose to opt-out tumours that had a large portion growing into the hila and mediastinum but kept tumours that were growing slightly into the hila and tumours with complex necrosis. One way to compare segmentation correlation between a proposed new technology (AI) and ground truth (manual segmentation) is to calculate the intersection over union (IoU) or Dice's coefficient. Arguably, if the AI's segmentation has excellent correlation with the manual segmentation, then TLG should also have similar correlation. Optimally, a fully automatic AI would have correlation scores close to 1 (perfect overlap) for both IoU and Dice's coefficient. However, most lung tumours detected with PET-CT are relatively small and might skew these correlative analyses. In reality, manual segmentations by expert readers do not have perfect overlap.

In Paper IV, AI was used to improve the PET image as a post-processing step after reconstruction. Evaluating image quality visually can only be done with readers and with that comes subjectivity. In order to mitigate this, a training/calibration session

was held before individual scoring. Readers used the clinical reconstruction as baseline (score = 3) and scored the reconstructions on a scale from 1 to 5. A senior radiology resident was included as a reader, in contrast to two senior nuclear medicine physicians, but ranked the reconstructions similarly.

# Conclusions

## Paper I

AI measurement of $SUV_{max}$ is up to 5 times faster, has almost-perfect reproducibility and good agreement compared with manual methods.

## Paper II

DS classification can change depending on the choice of reconstruction method, specifically between the old and new EARL recommendations. The latter was more prone to classify patients as non-responders.

## Paper III

AI can be used to detect lung lesions with high sensitivity and has high negative predictive value on a patient basis.

## Paper IV

AI can enhance BSREM reconstructed [$^{18}$F]FDG PET examinations to reduce noise and increase contrast compared with standard images whilst keeping $SUV_{max/peak}$ stability and decreasing the COV. The improved image quality can potentially be used either to provide better images to physicians or to reduce acquisition time/administered activity.

# Future perspectives

Although feasibility of using AI to aid nuclear medicine physicians has been shown in multiple studies, implementation into major software suites is slow. The number of PET-CT exams is increasing yearly, but the amount of expert readers is not keeping up with the growth. Therefore, it is important that work is also progressing in other domains such as software development, laws and education to facilitate the use of AI in clinical practise.

New hardware and reconstruction methods will continuously improve the spatial resolution of PET exams. When these technological upgrades have an impact on grading systems such as Deauville score, it would be interesting to see if these predict overall survivability with greater accuracy in larger cohorts. Quantification of underlying parameters should also be standardised and used more often instead of visual interpretation.

Further investigation into AI-models that are capable of segmenting complex tumours with growth into the mediastinum or with necrosis is needed. Calculation of TLG in these tumours remains a difficult and time-consuming task if done manually. An automatic solution could provide individual prognosis in the future. Guidelines and prognostic biomarkers like TLG which rely on quantitative rather than visual measurement will likely have a greater chance at adoption by image reading experts.

The feasibility of enhancing PET images with AI in different ways, enabling lower injected dose of radioactive tracers is positive and may lead to wider application of the modality. But the progress must be accompanied with validation studies that compare the sensitivity and specificity of these novel methods with the current standard.

The state of AI research today in imaging is focused on solving specific problems or tasks aimed at specific datasets. A general AI which is able to independently formulate research questions, analyse and solve these on its own is commonly referred to as technological singularity – such technology would not only make researchers and imaging specialists obsolete but also give rise to existential risks for humankind.

# Errata

## Paper I

The authors have discovered that the wrong abstract was published and that there was a mistake regarding the limits of agreement in Figure 2 (in the manuscript of Paper I). Please find the correct abstract below and the correct limits of agreement in Figures 7-9 in this thesis. The correct results further strengthen the conclusion.

*Correct abstract*

Background: Changes in maximum standardised uptake values ($SUV_{max}$) between serial PET-CT studies are used to determine disease progression or regression in oncologic patients. To measure these changes manually can be time consuming in a clinical routine. A semi-automatic method for calculation of $SUV_{max}$ in serial PET-CT studies was developed and compared to a conventional manual method. The semi-automatic method first aligns the serial PET-CT studies based on the CT images. Thereafter, the reader selects an abnormal lesion in one of the PET studies. After this manual step, the program automatically detects the corresponding lesion in the other PET study, segments the two lesions and calculates the $SUV_{max}$ in both studies as well as the difference between the $SUV_{max}$ values. The results of the semi-automatic analysis were compared to that of a manual $SUV_{max}$ analysis using a Philips PET-CT workstation. Three readers did the $SUV_{max}$ readings in both methods. Sixteen patients with lung cancer or lymphoma who had undergone two PET-CT studies were included. There was a total of 26 lesions.

Results: Bland-Altman's 95% limits of agreement were 0.46 to -1.88; 0.77 to -1.92 and 0.23 to -1.31 for each reader. Manual and semi-automatic method agreed in all cases whether $SUV_{max}$ had increased or decreased between serial studies. The average time to measure $SUV_{max}$ changes in two serial PET-CT examinations was four to five times longer for the manual method compared to the semi-automatic method for all readers (reader 1: 53.7 vs. 10.5 s; reader 2: 27.3 vs. 6.9 s; reader 3: 47.5 vs. 9.5 s; p < 0.001 for all).

Conclusions: Good agreement was shown in assessment of $SUV_{max}$ changes between manual and semi-automatic method. The semi-automatic analysis was four to five times faster to perform than the manual analysis. These findings show the feasibility of using semi-automatic methods for calculation of $SUV_{max}$ in clinical routine and encourage further development of programs using this type of methods.

*Correct Bland-Altman's limits of agreement*

Fig. 2A: Reader 1, 95% limits of agreement 0.46 to -1.88.

Fig. 2B: Reader 2, 95% limits of agreement 0.77 to -1.92.

Fig. 2C: Reader 3, 95% limits of agreement 0.23 to -1.31.

## Paper II

In Paper II, it was stated that patients were recruited retrospectively, which is incorrect – the study was done prospectively.

# Acknowledgements

# References

1.  Fukushima, K., *Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.* Biol Cybern, 1980. **36**(4): p. 193-202.
2.  Le Cun, Y., *Handwritten Digit Recognition with a Back-Propagation Network.* Proceedings of the Advances in Neural Information Processing Systems, 1989: p. 396–404.
3.  Hong, X., et al., *Enhancing the Image Quality via Transferred Deep Residual Learning of Coarse PET Sinograms.* IEEE Trans Med Imaging, 2018. **37**(10): p. 2322-2332.
4.  Song, T.A., et al., *Super-Resolution PET Imaging Using Convolutional Neural Networks.* IEEE Trans Comput Imaging, 2020. **6**: p. 518-528.
5.  Minarik, D., O. Enqvist, and E. Tragardh, *Denoising of Scintillation Camera Images using a Deep Convolutional Neural Network: A Monte Carlo Simulation Approach.* J Nucl Med, 2019.
6.  Lindgren Belal, S., et al., *Deep learning for segmentation of 49 selected bones in CT scans: First step in automated PET/CT-based 3D quantification of skeletal metastases.* Eur J Radiol, 2019. **113**: p. 89-95.
7.  Lindgren Belal, S., et al., *3D skeletal uptake of 18F sodium fluoride in PET/CT images is associated with overall survival in patients with prostate cancer.* EJNMMI Res, 2017. **7**(1): p. 15.
8.  Sadik, M., et al., *Automated quantification of reference levels in liver and mediastinal blood pool for the Deauville therapy response classification using FDG-PET/CT in Hodgkin and non-Hodgkin lymphomas.* Clin Physiol Funct Imaging, 2019. **39**(1): p. 78-84.
9.  Weisman, A.J., et al., *Comparison of 11 automated PET segmentation methods in lymphoma.* Phys Med Biol, 2020.
10. Kawauchi, K., et al., *A convolutional neural network-based system to classify patients using FDG PET/CT examinations.* BMC Cancer, 2020. **20**(1): p. 227.
11. Sibille, L., et al., *(18)F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks.* Radiology, 2020. **294**(2): p. 445-452.
12. Baek, S., et al., *Deep segmentation networks predict survival of non-small cell lung cancer.* Sci Rep, 2019. **9**(1): p. 17286.
13. Polymeri, E., et al., *Deep learning-based quantification of PET/CT prostate gland uptake: association with overall survival.* Clin Physiol Funct Imaging, 2019.

14.    Borrelli, P., et al., *Artificial Intelligence-based detection of lymph node metastases by PET/CT predicts prostate cancer-specific survival.* Clin Physiol Funct Imaging, 2020.

15.    Hosny, A., et al., *Artificial intelligence in radiology.* Nat Rev Cancer, 2018. **18**(8): p. 500-510.

16.    Czernin, J., *Clinical applications of FDG-PET in oncology.* Acta Med Austriaca, 2002. **29**(5): p. 162-70.

17.    Chen, H., et al., *Low-dose CT via convolutional neural network.* Biomed Opt Express, 2017. **8**(2): p. 679-694.

18.    Xiang, L., et al., *Deep Auto-context Convolutional Neural Networks for Standard-Dose PET Image Estimation from Low-Dose PET/MRI.* Neurocomputing, 2017. **267**: p. 406-416.

19.    *ACR Select - National Decision Support Company.* [cited 2020 15/10]; Available from: http://nationaldecisionsupport.com/acrselect/.

20.    *ESR iGuide | European Society of Radiology.* [cited 2020 15/10]; Available from: https://www.myesr.org/esriguide.

21.    van Assen, M., S.J. Lee, and C.N. De Cecco, *Artificial intelligence from A to Z: From neural network to legal framework.* Eur J Radiol, 2020. **129**: p. 109083.

22.    Coates, A., et al., *Deep learning with COTS HPC systems*, in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. 2013, JMLR.org: Atlanta, GA, USA. p. III–1337–III–1345.

23.    Forsting, M., *Machine Learning Will Change Medicine.* J Nucl Med, 2017. **58**(3): p. 357-358.

24.    Pesapane, F., et al., *Myths and facts about artificial intelligence: why machine- and deep-learning will not replace interventional radiologists.* Med Oncol, 2020. **37**(5): p. 40.

25.    *Evidence- based indications for the use of PET-CT in the United Kingdom*. 2016, The Royal College of Radiologists.

26.    Boellaard, R., et al., *FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0.* Eur J Nucl Med Mol Imaging, 2015. **42**(2): p. 328-54.

27.    Hsu, D.F.C., et al., *Studies of a Next-Generation Silicon-Photomultiplier-Based Time-of-Flight PET/CT System.* J Nucl Med, 2017. **58**(9): p. 1511-1518.

28.    Wagatsuma, K., et al., *Comparison between new-generation SiPM-based and conventional PMT-based TOF-PET/CT.* Phys Med, 2017. **42**: p. 203-210.

29.    Oddstig, J., et al., *Comparison of conventional and Si-photomultiplier-based PET systems for image quality and diagnostic performance.* BMC Med Imaging, 2019. **19**(1): p. 81.

30.    Tong, S., A.M. Alessio, and P.E. Kinahan, *Image reconstruction for PET/CT scanners: past achievements and future challenges.* Imaging Med, 2010. **2**(5): p. 529-545.

31. Iriarte, A., et al., *System models for PET statistical iterative reconstruction: A review.* Comput Med Imaging Graph, 2016. **48**: p. 30-48.

32. Hudson, H.M. and R.S. Larkin, *Accelerated image reconstruction using ordered subsets of projection data.* IEEE Trans Med Imaging, 1994. **13**(4): p. 601-9.

33. Ross, S., *Q.Clear, GE Healthcare.* White paper, 2014.

34. Economou Lundeberg, J., et al., *Comparison between silicon photomultiplier-based and conventional PET/CT in patients with suspected lung cancer-a pilot study.* EJNMMI Res, 2019. **9**(1): p. 35.

35. Lindstrom, E., et al., *Evaluation of Penalized-Likelihood Estimation Reconstruction on a Digital Time-of-Flight PET/CT Scanner for (18)F-FDG Whole-Body Examinations.* J Nucl Med, 2018. **59**(7): p. 1152-1158.

36. Teoh, E.J., et al., *Phantom and Clinical Evaluation of the Bayesian Penalized Likelihood Reconstruction Algorithm Q.Clear on an LYSO PET/CT System.* J Nucl Med, 2015. **56**(9): p. 1447-52.

37. Chen, K.T., et al., *Ultra-Low-Dose (18)F-Florbetaben Amyloid PET Imaging Using Deep Learning with Multi-Contrast MRI Inputs.* Radiology, 2019. **290**(3): p. 649-656.

38. Boykov, Y.Y. and M. Jolly. *Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images*. in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. 2001.

39. Ronneberger, O., Fischer, P., Brox, T., *U-Net: Convolutional Networks for Biomedical Image Segmentation.* arXiv:1505.04597, 2015.

40. Boellaard, R., et al., *FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0.* Eur J Nucl Med Mol Imaging, 2010. **37**(1): p. 181-200.

41. Kaalep, A., et al., *Feasibility of state of the art PET/CT systems performance harmonisation.* Eur J Nucl Med Mol Imaging, 2018. **45**(8): p. 1344-1361.

42. *EARL: Accreditation specifications*. [cited 2020 15/10]; Available from: http://earl.eanm.org/cms/website.php?id=/en/projects/fdg_pet_ct_accreditation/accreditation_specifications.htm.

43. Barrington, S.F. and M. Meignan, *Time to Prepare for Risk Adaptation in Lymphoma by Standardizing Measurement of Metabolic Tumor Burden.* J Nucl Med, 2019. **60**(8): p. 1096-1102.

44. Sollini, M., et al., *PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology.* Sci Rep, 2017. **7**(1): p. 358.

45. Ha, S., et al., *Radiomics in Oncological PET/CT: a Methodological Overview.* Nucl Med Mol Imaging, 2019. **53**(1): p. 14-29.

46. Danckert, B. *NORDCAN: Cancer Incidence, Mortality, Prevalence, and Survival in the Nordic Countries, Version 8.2 (26.03.2019). Association of the Nordic Cancer Registries. Danish Cancer Society.* [cited 2020 15/10]; Available from: http://www.ancr.nu.

47.	*SEER*Explorer: An interactive website for SEER cancer statistics [Internet]. Surveillance Research Program, National Cancer Institute.* [cited 2020 15/10]; Available from: https://seer.cancer.gov/explorer/.

48.	Meignan, M., et al., *Report on the First International Workshop on Interim-PET-Scan in Lymphoma.* Leuk Lymphoma, 2009. **50**(8): p. 1257-60.

49.	Radford, J., et al., *Results of a trial of PET-directed therapy for early-stage Hodgkin's lymphoma.* N Engl J Med, 2015. **372**(17): p. 1598-607.

50.	Borchmann, P., et al., *PET-guided treatment in patients with advanced-stage Hodgkin's lymphoma (HD18): final results of an open-label, international, randomised phase 3 trial by the German Hodgkin Study Group.* Lancet, 2018. **390**(10114): p. 2790-2802.

51.	Kobe, C., et al., *Outcome-based interpretation of early interim PET in advanced-stage Hodgkin lymphoma.* Blood, 2018. **132**(21): p. 2273-2279.

52.	Sauter, C.S., et al., *Prognostic value of FDG-PET prior to autologous stem cell transplantation for relapsed and refractory diffuse large B-cell lymphoma.* Blood, 2015. **125**(16): p. 2579-81.

53.	Alcantara, M., et al., *PET/CT before autologous stem cell transplantation predicts outcome in refractory/relapsed follicular lymphoma.* Eur J Nucl Med Mol Imaging, 2015. **42**(2): p. 215-21.

54.	Kim, S.J., et al., *Risk stratification on the basis of Deauville score on PET-CT and the presence of Epstein-Barr virus DNA after completion of primary treatment for extranodal natural killer/T-cell lymphoma, nasal type: a multicentre, retrospective analysis.* Lancet Haematol, 2015. **2**(2): p. e66-74.

55.	Collaborators, G.B.D.C.o.D., *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): p. 1736-1788.

56.	*Smoking Cessation: A Report of the Surgeon General — executive summary.* 2020, US Department of Health and Human Services, Rockville, MD.

57.	Pieterman, R.M., et al., *Preoperative staging of non-small-cell lung cancer with positron-emission tomography.* N Engl J Med, 2000. **343**(4): p. 254-61.

58.	van Tinteren, H., et al., *Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: the PLUS multicentre randomised trial.* Lancet, 2002. **359**(9315): p. 1388-93.

59.	Pellegrino, S., et al., *Total metabolic tumor volume by 18F-FDG PET/CT for the prediction of outcome in patients with non-small cell lung cancer.* Ann Nucl Med, 2019. **33**(12): p. 937-944.

60.	Hyun, S.H., et al., *Volume-based parameters of (18)F-fluorodeoxyglucose positron emission tomography/computed tomography improve outcome*

*prediction in early-stage non-small cell lung cancer after surgical resection.* Ann Surg, 2013. **257**(2): p. 364-70.

61.     Jones, T. and D. Townsend, *History and future technical innovation in positron emission tomography.* J Med Imaging (Bellingham), 2017. **4**(1): p. 011013.

62.     *Isotopstatistik för nukleärmedicinsk verksamhet, Strålsäkerhetsmyndigheten.* [cited 2020 15/10]; Available from: https://dosreg.ssm.se/Isotopstatistik/RegistreringPublik.

63.     Moon, S.H., S.H. Hyun, and J.Y. Choi, *Prognostic significance of volume-based PET parameters in cancer patients.* Korean J Radiol, 2013. **14**(1): p. 1-12.

64.     Vensby, P.H., et al., *The value of FDG PET/CT for follow-up of patients with melanoma: a retrospective analysis.* Am J Nucl Med Mol Imaging, 2017. **7**(6): p. 255-262.

65.     Mena, E., et al., *18F-FDG PET/CT and Melanoma: Value of Fourth and Subsequent Posttherapy Follow-up Scans for Patient Management.* Clin Nucl Med, 2016. **41**(9): p. e403-9.

66.     Vikram, R. and R.B. Iyer, *PET/CT imaging in the diagnosis, staging, and follow-up of colorectal cancer.* Cancer Imaging, 2008. **8 Spec No A**: p. S46-51.

67.     Marcus, C., et al., *18F-FDG PET/CT and Colorectal Cancer: Value of Fourth and Subsequent Posttherapy Follow-up Scans for Patient Management.* J Nucl Med, 2015. **56**(7): p. 989-94.

68.     Mahmud, A., R. Poon, and D. Jonker, *PET imaging in anal canal cancer: a systematic review and meta-analysis.* Br J Radiol, 2017. **90**(1080): p. 20170370.

69.     Tragardh, E., et al., *RECOMIA-a cloud-based platform for artificial intelligence research in nuclear medicine and radiology.* EJNMMI Phys, 2020. **7**(1): p. 51.

70.     Zhang, K., et al., *Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising.* IEEE Trans Image Process, 2017. **26**(7): p. 3142-3155.

71.     Nair, A., et al., *Variable radiological lung nodule evaluation leads to divergent management recommendations.* Eur Respir J, 2018. **52**(6).

72.     Tanner, N.T., et al., *Physician Assessment of Pretest Probability of Malignancy and Adherence With Guidelines for Pulmonary Nodule Evaluation.* Chest, 2017. **152**(2): p. 263-270.

73.     Moore, G.E., *Cramming more components onto integrated circuits.* Electronics, 1965. **Volume 38**(8).

74.     Barrington, S.F. and R. Kluge, *FDG PET for therapy monitoring in Hodgkin and non-Hodgkin lymphomas.* Eur J Nucl Med Mol Imaging, 2017. **44**(Suppl 1): p. 97-110.

75.     Nekolla, S.G., et al., *Evaluation of the novel myocardial perfusion positron-emission tomography tracer 18F-BMS-747158-02: comparison*

to 13N-ammonia and validation with microspheres in a pig model. Circulation, 2009. **119**(17): p. 2333-42.

76.     Kaalep, A., et al., *Quantitative implications of the updated EARL 2019 PET-CT performance standards.* EJNMMI Phys, 2019. **6**(1): p. 28.

77.     Jacene, H.A., et al., *Assessment of interobserver reproducibility in quantitative 18F-FDG PET and CT measurements of tumor response to therapy.* J Nucl Med, 2009. **50**(11): p. 1760-9.

78.     Yan, J., et al., *Automated matching and segmentation of lymphoma on serial CT examinations.* Med Phys, 2007. **34**(1): p. 55-62.

79.     Enilorac, B., et al., *Does PET Reconstruction Method Affect Deauville Score in Lymphoma Patients?* J Nucl Med, 2018. **59**(7): p. 1049-1055.

80.     Wyrzykowski, M., et al., *Impact of the Q.Clear reconstruction algorithm on the interpretation of PET/CT images in patients with lymphoma.* EJNMMI Res, 2020. **10**(1): p. 99.

# Artificial intelligence in PET-CT

John Ly is a board-certified radiologist, currently working at Centralsjukhuset Kristianstad in Skåne, Sweden. In this thesis he has investigated how artificial intelligence can add value to PET-CT research and clinical applications. It is believed by experts that this transformative technology will continuously improve medical imaging in many facets such as radiation dose, image reconstruction and quality, detection, classification, monitoring, reporting, workflows and even interventional radiology. The main conclusion is that the artificial intelligence algorithms used in this thesis were able to function as a complement to and increase the quality of PET-CT examinations.

## FACULTY OF MEDICINE

LUND UNIVERSITY