Mri-Based Radiomics in Breast Cancer

MRI-BASED RADIOMICS IN BREAST CANCER

Optimization and Prediction

Renée W. Y. Granzier

Cover design & layout: Publiss | www.publiss.nl
Print: Ridderprint | www.ridderprint.nl
ISBN: 978-94-6416-752-8

The research presented in this thesis was conducted at GROW-School for Oncology and Developmental Biology, Department of Surgery, Maastricht University.

Parts of the research described in this thesis were financially supported by Kankeronderzoeksfonds Limburg and by a grant from KWF Kankerbestrijding (Project number 12085/2018-2).

MRI-BASED RADIOMICS IN BREAST CANCER

Optimization and Prediction

Proefschrift

Ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. Dr. Pamela Habibović
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 22 September 2022 om 13:00 uur

door

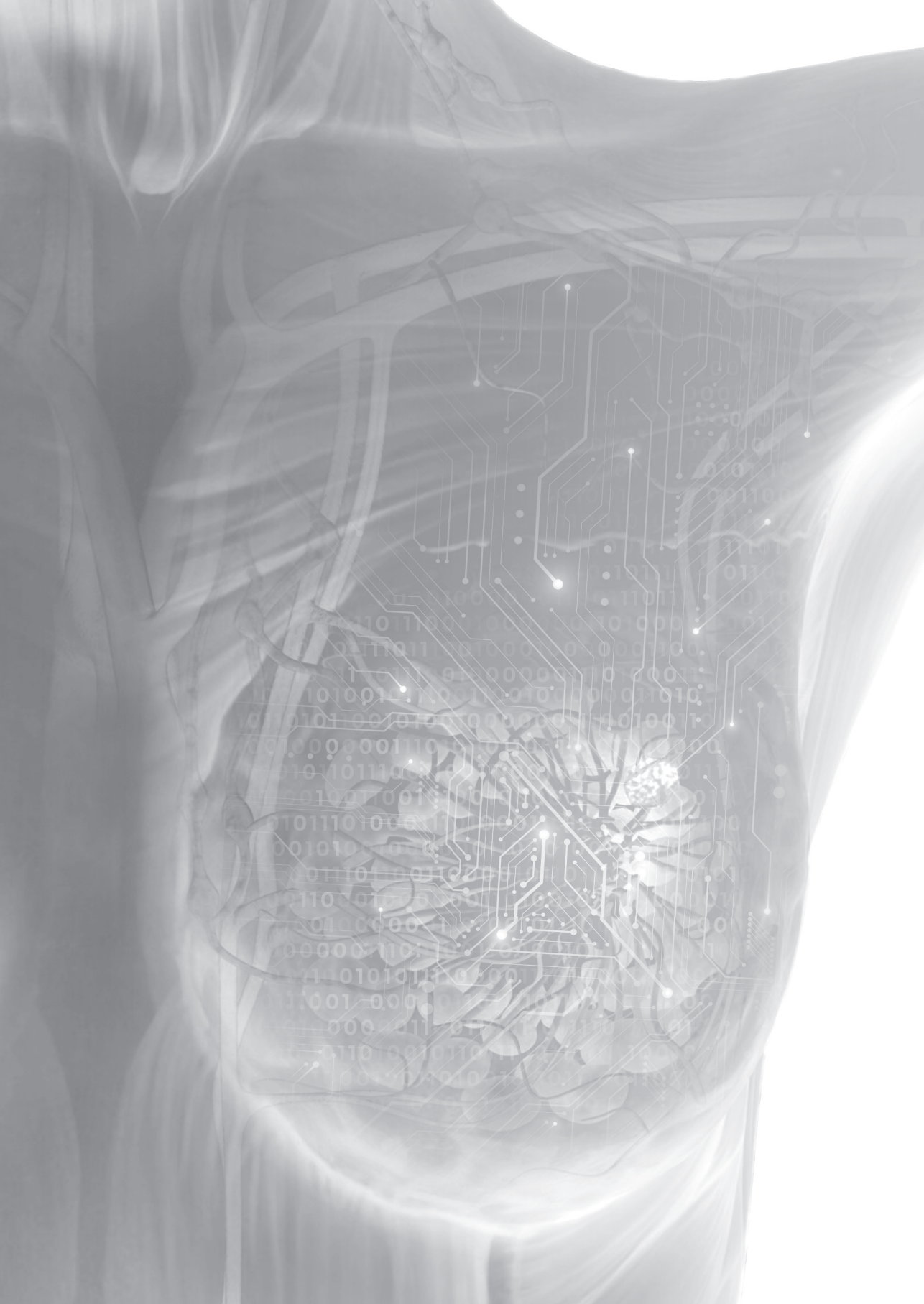
Renée W.Y. Granzier

Promotor
Prof. dr. M. L. Smidt

Co-promotores
Dr. H.C. Woodruff
Dr. M.B.I. Lobbes

Beoordelingscomissie
Prof. dr. V.C.G. Tjan-Heijnen (voorzitter)
Prof. dr. W.J. Niessen
Prof. dr. R.M. Pijnappel
dr. L.Y.L. Wee

Table of Contents



CHAPTER 1

Introduction and thesis outline

With more than 17,000 new breast cancer diagnoses in 2019, breast cancer is the most common cancer among women in the Netherlands. Currently, during their lifetime, one in seven women will develop breast cancer [1]. As of 2020, breast cancer is the most common cancer worldwide, with 2.26 million new cases of breast cancer [2]. Breast cancer treatment is based on three pillars, namely surgery, systemic therapy, and radiotherapy.

Where surgical treatment has become less invasive, treatment options within systemic therapy and radiotherapy have expanded and improved. Whereas in the 1980s the standard surgical treatment consisted of a modified radical mastectomy, in which the entire breast and all axillary lymph nodes were removed, breast-conserving treatment with the removal of the sentinel lymph node for lymph node staging, is nowadays mostly performed [3,4]. This is a less invasive treatment that has been proven to be oncologically safe [5,6]. These treatment changes, as well as advances in the screening and diagnosis, resulted in a considerable improvement in overall survival rates, with 5-year survival rates of women with invasive breast cancer increasing from 55% to 88% between 1961 and 2017 [1].

Despite these advancements, there are still downsides. All surgical procedures can cause co-morbidities such as numbness and tingling of the skin, shoulder stiffness, seroma, and lymphedema, which can have a significant impact on quality of life [7]. Systemic therapy is not only toxic to the cancer cells but also affects the healthy cells of our body, causing fertility problems, nausea, vomiting, alopecia, hypersensitive skin reactions, mucositis, and neurotoxicity [8]. Though these side effects again have an impact on quality of life, fortunately, most patients will (partly) recover over time after cessation of the systemic therapy. Side effects of the radiotherapy are mainly skin related, including skin redness, blistering, and itching during and after radiotherapy. Over the long term, the irradiated skin may darken and thicken, with possible changes in breast size. In addition to the skin-related side effects, the heart and lung region are also exposed to radiation, although this is limited by hypofractionation and the breath-hold irradiation technique.

Systemic therapy & tumor response

The type of systemic therapy is determined according to specific patient and tumor characteristics. Systemic therapy can consist of hormonal therapy, chemotherapy, or immunotherapy, alone or in a combination. In addition, systemic therapy can be administered before surgical treatment (i.e., neoadjuvant systemic therapy) or afterward (i.e., adjuvant systemic therapy). The current trend is the increasing application of neoadjuvant systemic treatment (NST), as this has some advantages. Firstly, it is possible to monitor the *in vivo* tumor response and thereby facilitate the study of cancer biology. Secondly, NST might enable less invasive surgical intervention due to tumor shrinkage.

Not all tumors respond well to NST. Tumor response to NST ranges from complete disappearance of the tumor, so-called pathologic complete response (pCR), to pathologic partial response, non-response, or even progression of the disease. Tumors achieving pCR are associated with improved disease-free survival and overall survival compared to tumors that do not achieve pCR [9]. Approximately 30% of all breast cancer patients treated with NST achieve a pCR [10,11].

Previous research on tumor response to NST has provided insight into various patient and tumor characteristics and their associated tumor response. For example, the breast cancer subtype is a determinant of tumor response to NST. Tumors with a human epidermal growth factor receptor 2 (HER2) positive status and negative estrogen (ER) and progesterone (PR) receptor have a high potential for a pCR after treatment with targeted therapy pertuzumab and trastuzumab [12].

Breast tumor size is another important predictor of tumor response, with small tumors achieving significantly higher pCR rates than larger tumors [13]. Studies have also shown that tumor response patterns are an important predictor in determining tumor response to NST [14]. For example, the study of Goorts et al. showed that MRI-based response patterns halfway NST predicted pCR more accurately than MRI-based response patterns after NST, with 83% of accurate pCR predictions based on MRI exams scanned halfway NST and 41% of accurate pCR predictions based on MRI exams scanned after NST. Furthermore, tumors showing concentric shrinking showed higher pCR rates compared to tumors that crumbled.

However, variation in tumor response also exists between nearly identical tumors (i.e., two patients with breast tumors of the same subtype, and tumor size) treated with the same NST. These differences in tumor response are likely caused by unknown factors that are not (yet) observable by both the radiologist and the pathologist.

Imaging

Assessment of the breast tumor and monitoring the *in vivo* tumor response before, during, and after NST is performed by medical imaging. Imaging modalities that can be considered for assessment and monitoring of the breast tumor response include (full-field digital) mammography, contrast-enhanced mammography (CEM), ultrasound, positron emission tomography-computed tomography (PET-CT), and magnetic resonance imaging (MRI).

Assessment of breast tumor extent, including multifocality and multicentricity, as well as the assessment of contralateral breast tumors, is usually performed through MRI exams [15,16]. Sensitivity values for the diagnosis of breast cancer

using MRI exams range between 85% and 100% [17,18]. More recent studies also showed that contrast-enhanced MRI reaches the highest accuracy for breast tumor response monitoring [19,20] (Figure 1). Although breast MRI is the most accurate imaging modality for response monitoring, the use of breast MRI, even in combination with clinical patient and tumor characteristics, is not (yet) sufficiently accurate to provide information to effect changes in clinical treatment, as both overestimation and underestimation of tumor response are observed. If pCR could be predicted accurately prior to surgery, surgery, and/or radiotherapy, could potentially be omitted. Conversely, if tumors can be predicted not to respond to NST even before the start of NST or perhaps halfway through NST, toxic treatment can be discontinued and the patient can immediately proceed to surgical treatment with or without radiotherapy or choose to initiate a different treatment.

Unraveling the often heterogeneous nature of breast tumors may bring us one step closer to an accurate pCR prediction. Since an MRI contains much more information than a radiologist can perceive with the naked eye, a further, more quantitative, in-depth analysis of the MRI seems necessary. To do this, radiomics analysis may well be the missing piece of the puzzle.

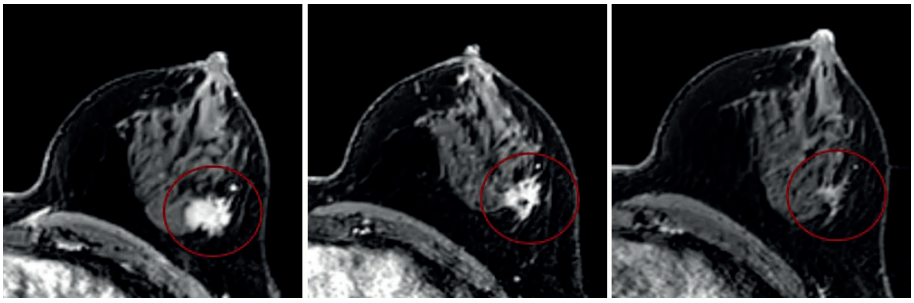


Figure 1. Contrast-enhanced, T1-weighted axial image of the left breast, in which an ill-defined, irregular mass is present (circle), before (A), during (B) and after (C) neoadjuvant systemic therapy, showing heterogeneous enhancement. Over time, a decrease in tumor size reduction and a decrease in signal intensity are visible.

Radiomics

The emergence of personalized medicine has gained momentum with the introduction of artificial intelligence into oncology care. Radiomics, an artificial intelligence workflow, has been used in recent years to analyze medical imaging for disease detection and characterization, as well as to help with diagnosis and predictions [21,22].

Radiomics is an image analysis method that extracts large amounts of quantitative handcrafted data, called features, from a defined region of interest (ROI) within a

medical image. With the advent of radiomics, medical images are not only visually analyzed for tumor characterization, but also analyzed in a quantitative way. The application of radiomics for oncological issues has expanded enormously in the latest years, due to rapid medical imaging technical developments, better image archiving, the increase in the use of medical imaging in clinics, and machine learning technique developments. The quantitative radiomics features extracted from medical images include shape, intensity, texture, and fractal features, which can be extracted from both the original or filtered image, and all contain image-derived information regarding the ROI [23].

In the current diagnosis of breast cancer, a tumor tissue biopsy is used, which apart from being invasive, represents only a small part of the tumor. Using radiomics, information is extracted from the complete ROI (usually the tumor in the application of radiomics for oncological problems) in a non-invasive manner, which is a great advantage in the knowledge that breast tumors have a heterogeneous nature.

The radiomics workflow consists of several steps including, image acquisition, image pre-processing, ROI segmentation, feature extraction, feature selection, and feature analysis (e.g. model development) (Figure 2).

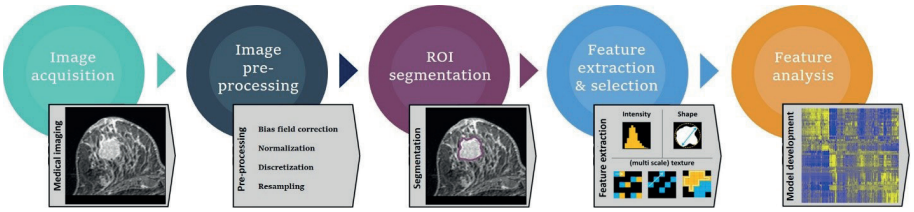


Figure 2. Radiomics workflow. Abbreviations: ROI, region of interest.

Image acquisition is the first and necessary step to start the radiomics analysis. Specifically, for breast cancer imaging, this includes, CT, MRI, PET, mammography, ultrasound, and CEM. However, the use of imaging modalities where gray level values correspond to measurable quantities (e.g. Hounsfield Units in CT imaging) for radiomics analyses is preferable compared to imaging modalities where image intensities are hard to quantify as gray levels can take on variable values, as in, MRI. The latter is the modality of interest in this thesis.

After collecting the medical images, it is important that the images are preprocessed to obtain more comparable images before features can be extracted. This is because images are often collected retrospectively from different scanners and over a long period of time, leading to heterogeneity in image acquisition and

reconstruction parameters. To extract radiomics features it is necessary to define an ROI within the image. In the application of radiomics in oncology, the tumor is usually chosen as the ROI. In this thesis, all segmentations were performed in 3D, and multiple regions were chosen as ROI, including breast tumors, axillary lymph nodes, and the complete breast.

Once the ROI is defined on the preprocessed images it is possible to extract over a thousand features per ROI. To develop meaningful and robust radiomics models, it is necessary to perform feature selection. It is common for many characteristics to be correlated and/or unrelated to the outcome, often due to the sheer number of radiomic features. Including these features in the model can lead to the introduction of noise. Furthermore, features sensitive to changes induced by scanning variabilities (i.e. different scanners or varying acquisition and reconstruction parameters) can affect feature values and should be removed. There is also a risk of overfitting if more features than events are included in the model. This can often be identified by the error rate that is larger for the test and/or validation dataset compared to the error rate for the training dataset. In such cases, the developed model is not able to generalize the results or fit well to new, unseen data.

Now with a set of preselected features, modeling can start. There is a wide variety of machine learning techniques that can be applied to develop prognostic or predictive models. The random forest model is a supervised learning method, utilizing an ensemble of decision trees, which is often used for classification purposes and the method of choice in this thesis. Random forest is based on the fact that the combination of different machine learning techniques leads to an improved result.

To conclude, this thesis describes the application of radiomics, a quantitative image analysis method, within magnetic resonance imaging (MRI) of breast cancer patients. It investigated the predictive power of MRI-based radiomics in the treatment of breast cancer patients and worked on optimization of this specific area of research.

Thesis outline

This thesis consists of two parts that are preceded by an explanation of the radiomics workflow in chapter two. This provides an extensive explanation of the radiomics workflow supplemented with a general picture of the current use of radiomics in the oncological field, including an overview of the current limitations.

Part 1 of this thesis focuses on the use of MRI-based radiomics for prediction analyses in the treatment of breast cancer patients. In chapter three, an overview and quality assessment of articles published until May 2019 on the prediction of tumor response to NST using MRI-based radiomics in breast cancer patients is provided in a narrative systematic review. In chapter four, the potential of MRI-based radiomics for the prediction of tumor response to NST using retrospectively collected pretreatment MRI exams of 320 breast cancer patients from two different hospitals is investigated. In chapter five, a prospectively collected data cohort, consisting of 90 breast cancer patients is used to investigate whether axillary nodal metastasis could be predicted using MRI-based dedicated axillary T2-weighted (T2W) imaging.

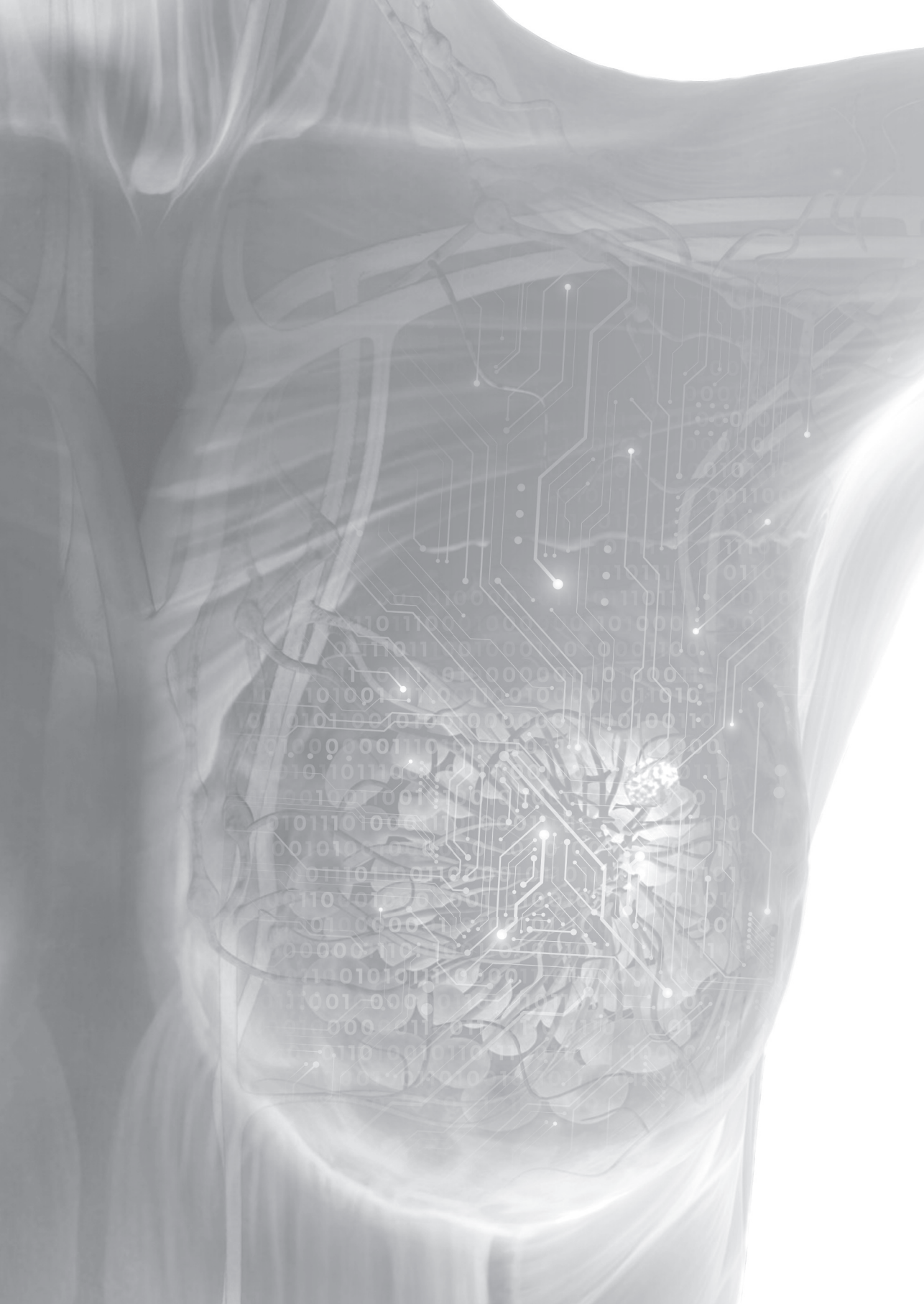
The second part of this thesis focuses on specific steps of the radiomics workflow intending to improve these steps and thereby obtain a more applicable and generalizable radiomics workflow. In chapter six, the repeatability of MRI radiomics features obtained from test-retest data from eleven healthy volunteers is studied to determine repeatable features. In chapter seven, the reproducibility of MRI breast segmentations, performed by four different observers, and their influence on feature values, extracted by two commonly used radiomics software, is reported.

Chapter eight contains a general discussion of the thesis and further elaboration of future perspectives on MRI-based radiomics research to improve personalized medicine in breast cancer patients. Finally, a summary of the thesis can be found in chapter nine.

References

1. Nederlandse Kankerregistratie IKNL. [Available from: https://iknl.nl/nkr-cijfers]. 2021.
2. World Health Organization. [Available from: https://www.who.int/news-room/fact-sheets/detail/cancer] 2021
3. Madden JL, Kandalaft S, Bourque RA. Modified radical mastectomy. Annals of surgery. 1972;175(5):624-34.
4. van Dongen JA, Voogd AC, Fentiman IS, Legrand C, Sylvester RJ, Tong D, et al. Long-term results of a randomized trial comparing breast-conserving therapy with mastectomy: European Organization for Research and Treatment of Cancer 10801 trial. Journal of the National Cancer Institute. 2000;92(14):1143-50.
5. Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A, et al. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. The New England journal of medicine. 2002;347(16):1227-32.
6. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. The New England journal of medicine. 2002;347(16):1233-41.
7. Fu MR, Axelrod D, Guth AA, Cleland CM, Ryan CE, Weaver KR, et al. Comorbidities and Quality of Life among Breast Cancer Survivors: A Prospective Study. J Pers Med. 2015;5(3):229-42.
8. Gudgeon A. Side-effects of systemic therapy for the management of breast cancer. S Afr Med J. 2014;104(5):381.
9. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet. 2014;384(9938):164-72.
10. Haque W, Verma V, Hatch S, Suzanne Klimberg V, Brian Butler E, Teh BS. Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy. Breast cancer research and treatment. 2018;170(3):559-67.
11. von Minckwitz G, Untch M, Blohmer JU, Costa SD, Eidtmann H, Fasching PA, et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol. 2012;30(15):1796-804.
12. Symmans WF, Wei C, Gould R, Yu X, Zhang Y, Liu M, et al. Long-Term Prognostic Risk After Neoadjuvant Chemotherapy Associated With Residual Cancer Burden and Breast Cancer Subtype. J Clin Oncol. 2017;35(10):1049-60.
13. Goorts B, van Nijnatten TJ, de Munck L, Moossdorff M, Heuts EM, de Boer M, et al. Clinical tumor stage is the most important predictor of pathological complete response rate after neoadjuvant chemotherapy in breast cancer patients. Breast cancer research and treatment. 2017;163(1):83-91.
14. Goorts B, Dreuning KMA, Houwers JB, Kooreman LFS, Boerma EG, Mann RM, et al. MRI-based response patterns during neoadjuvant chemotherapy can predict pathological (complete) response in patients with breast cancer. Breast Cancer Res. 2018;20(1):34.
15. Van Goethem M, Schelfout K, Kersschot E, Colpaert C, Verslegers I, Biltjes I, et al. MR mammography is useful in the preoperative locoregional staging of breast carcinomas with extensive intraductal component. European journal of radiology. 2007;62(2):273-82.

16. Lobbes MB, Prevos R, Smidt M, Tjan-Heijnen VC, van Goethem M, Schipper R, et al. The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: a systematic review. Insights Imaging. 2013;4(2):163-75.
17. Heywang-Köbrunner SH, Beck R. Contrast-enhanced MRI of the breast. 2nd ed. Berlin ; New York: Springer; 1996. x, 229 p. p.
18. Kilic F, Ogul H, Bayraktutan U, Gumus H, Unal O, Kantarci M, et al. Diagnostic magnetic resonance imaging of the breast. Eurasian J Med. 2012;44(2):106-14.
19. Bouzon A, Acea B, Soler R, Iglesias A, Santiago P, Mosquera J, et al. Diagnostic accuracy of MRI to evaluate tumour response and residual tumour size after neoadjuvant chemotherapy in breast cancer patients. Radiol Oncol. 2016;50(1):73-9.
20. Weber JJ, Jochelson MS, Eaton A, Zabor EC, Barrio AV, Gemignani ML, et al. MRI and Prediction of Pathologic Complete Response in the Breast and Axilla after Neoadjuvant Chemotherapy for Breast Cancer: MRI and Pathologic Complete Response. J Am Coll Surg. 2017.
21. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer (Oxford, England : 1990). 2012;48(4):441-6.
22. Gillies R, Kinahan P, Hricak H. Radiomics: Images are more than pictures, they are data. Radiology. 2016;278(2):563-77.
23. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77(21):e104-e7.



CHAPTER 2

Radiomics: from qualitative to quantitative imaging

William Rogers, Sithin Thulasi Seetha, Turkey Refaee, Relinde I.Y. Lieverse, Renée W. Y. Granzier, Abdalla Ibrahim, Simon Keek, Sebastian Sanduleanu, Sergey Primakov, Manon Beuque, Damiënne Marcus, Alexander van der Wiel, Fadila Zerka, Cary Oberije, Janita E. van Timmeren, Henry C Woodruff, Philippe Lambin

Br J Radiol 2020; 93: 20190948

Abstract

Historically, medical imaging has been a qualitative or semi-quantitative modality. It is difficult to quantify what can be seen in an image, and to turn it into valuable predictive outcomes. As a result of advances in both computational hardware and machine learning algorithms, computers are making great strides in obtaining quantitative information from imaging and correlating it with outcomes. This opens a new "omics" field, radiomics, adding new input avenues for precision medicine, beyond genomics. Radiomics, in its two forms "handcrafted and deep", is an emerging field that translates medical images into quantitative data to yield biological information and enable radiologic phenotypic profiling for diagnosis, theragnosis, decision support, and monitoring. Within this review, we describe the steps of handcrafted radiomics, a multistage process in which features based on shape, pixel intensities, and texture are extracted from radiographs. The application of deep learning, the second arm of radiomics, and its place in the radiomics workflow is discussed, along with its advantages and disadvantages. To better illustrate the technologies being used, we provide real-world clinical applications of radiomics in oncology and other diseases, showcasing research on the applications of radiomics, as well as covering its limitations and its future direction towards precision medicine.

Introduction

Medical imaging technologies in healthcare have expanded remarkably from the discovery of X-Rays 124 years ago to the use of CT, MRI, and Positron Emission Tomography (PET), among others in modern-day clinical practice[1] (Figure 1). These tools have become an integral part in detection and diagnosis for many diseases due to several factors, including: the minimally invasive nature of imaging, rapid technological developments, lower costs compared to alternatives, the high information density of images, and the hardware can be used for multiple diseases and sites.[2,3]

Medical imaging in its infancy generated analogue images, which underwent subjective interpretation based on visual inspection and verbal communication. By the end of the 20th century, information technology has brought radiology to the digital world,[4] although the interpretation of radiographs remained mostly qualitative. Humans excel at recognising patterns through visual inspection, however, they are often lacking when performing complex quantitative assessments.[5,6] In the early 1960s, researchers started to focus on computerized quantitative analysis of medical data for aiding clinical diagnosis,[7-9] what later came to be known as Computer Aided Decision (CAD) systems. However, these systems were using a classical approach using statistical analysis and probability theories, and the volume of available data was low, so the results were often too inaccurate for clinical use. Later in the 1980s, further advances in theoretical computer science and digital imaging lead to the development of advanced machine learning and pattern recognition algorithms, which when integrated with CAD systems were able to generate clinically reliable results.[10,11]

In recent decades, simple quantitative image analysis (QIA) has been adopted by clinicians (*e.g.* RECIST[12]), and has been primarily focused on assisting qualitative observations.[13] For instance, CAD systems can be found in health care worldwide, aiding radiologists and clinicians in making diagnostic and theragnostic decisions.[14] One of the most typical applications of CAD systems is in recognizing abnormalities during cancer screening.[15] Notable contributions are in the area of lung and breast cancer research. For example, there are many CAD studies which focus on detecting and diagnosing lung nodules[16,17] (as benign or malignant) on CT and chest radiographs. Similarly, many such studies have been conducted in breast mammography images for highlighting microcalcifications,[18] architectural distortions, and the prediction of mass type.[19,20]

It is conceivable that the lack of quantitative information leads to increased follow-ups or invasive biopsies that would be deemed unnecessary given the unused information in medical images.[21] Even though there have been various developments in quantitative image analysis, traditionally radiologists are trained

to understand the behaviour of the underlying disease through visual inspection of radiographic images.[21] This partially explains why most of the developments in imaging technology are in optimising the visual representation of the generated images, with vendors competing to generate the highest quality images. With the exception of CT, with its semi-parametric calibrated Hounsfield Units, and some particular MRI sequences, individual voxel values do not correlate with the underlying biology without further calibration and modelling. Furthermore, qualitative analysis is not so dependent on reproducible voxel values, while machines on the other hand only process numerical values and rely on the standardisation of image acquisition and reconstruction to yield reproducible results. The lack of standardisation of medical images has been a major hurdle in the development of QIA in medical imaging.[22-25] However, in recent years, quantitative imaging is becoming more popular with the advent of, *e.g.*, quantitative fludeoxyglucose-PET[26,27] or quantitative MRI[28,29] for treatment response assessment.

The ubiquitous computer, vast amounts of data, and advanced algorithms have opened a new era in medical imaging. The high information density of images allows for many quantitative metrics since intricate pixel and voxel relationships can be captured by complex operations. Radiomics involves the process of extraction of quantifiable features from vast amounts of data that might correlate with the underlying biology or clinical outcomes using advanced machine learning analysis techniques.[30,31] Radiomics has two main arms, based on how imaging information is transformed into mineable data: handcrafted radiomics and deep learning. Handcrafted features are formulas mostly based on intensity histograms, shape attributes, and texture, that can be used to fingerprint phenotypical characteristics of the radiograph[32] while in deep learning a complex network "creates" its own features. Various statistical and machine learning models have been widely researched, and are envisioned to be complementary to best medical practice by aiding in making informed clinical decisions in both oncological and non-oncological diseases.[33-36]

Since the 1990s predictions were being made that genomics, spearheaded by the Human Genome Project, would completely transform therapeutic medicine, heralding precision medicine.[37] Precision medicine, also termed personalized medicine, originally referred to the view that incorporating genomic information in the clinical workflow will lead to marked improvements in the prediction, diagnosis, and treatment of diseases. Recently, the scope of precision medicine has expanded to incorporate inputs beyond the genome.[38] Radiomics and other "-omic" developments, such as metabolomics and proteomics, are contributing to this a paradigm shift in medicine, where the focus has changed from standard clinical protocols based on trial populations to a personalised treatment tailored not only to the disease and site but also the patient, further enabling precision medicine.

In this review, we provide a broad overview and update on the fast-growing field of quantitative imaging research, focusing on the two arms "handcrafted radiomics and deep learning" describing some of its caveats and giving examples of the budding clinical implementation, the stepping stones towards precision medicine.

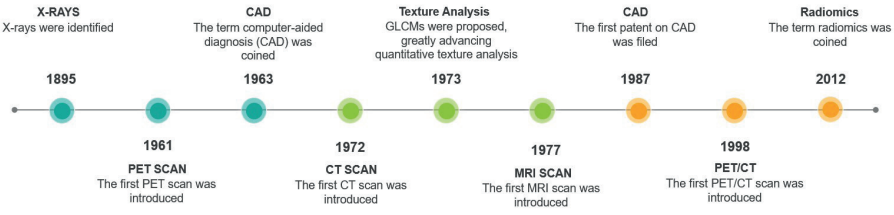


Figure 1. Timeline highlighting key developments in medical imaging. CAD, computer-aided diagnosis; GLCM, grey level co-occurring matrix; PET, positron emission tomography.

Radiomics: from feature extraction to correlation with outcomes

Performing feature extraction of textures in medical imaging is nothing new and in fact serious research had begun in the early 1980s at Kurt Rossmann Laboratories for Radiologic Image Research in the Department of Radiology at the University of Chicago to develop CAD systems for the detection of lung nodules as well as detection of clustered microcalcifications in mammograms[39,40]. The first CAD patent was filed all the way back in 1987 using a method of pixel thresholding and contiguous pixel area thresholding.[40]

The radiomic workflow begins with the medical image, which can be represented in two, three, or four dimensions.[32,41] Images contain quantitative data in the form of signals that are captured at different scales and variation across medical machines.[42,43] Normalisation techniques are used to distribute pixel intensities evenly across a dataset and within a standardized range.[42-44] Next, a region of interest (ROI) is defined so that only information related to the lesion can be extracted, and the useful information that can be extracted are called features. There are competing methods to extract features both in two-dimensional and three-dimensional. One such method is the manual segmentation of the lesion or the creation of a bounding box, as seen in Figure 2.[45,46] This can also be performed using automated segmentation algorithms. Methods for automated segmentation include deep learning architectures such as U-Net, or semi-automatic methods like click-and-grow algorithms.[45,46]

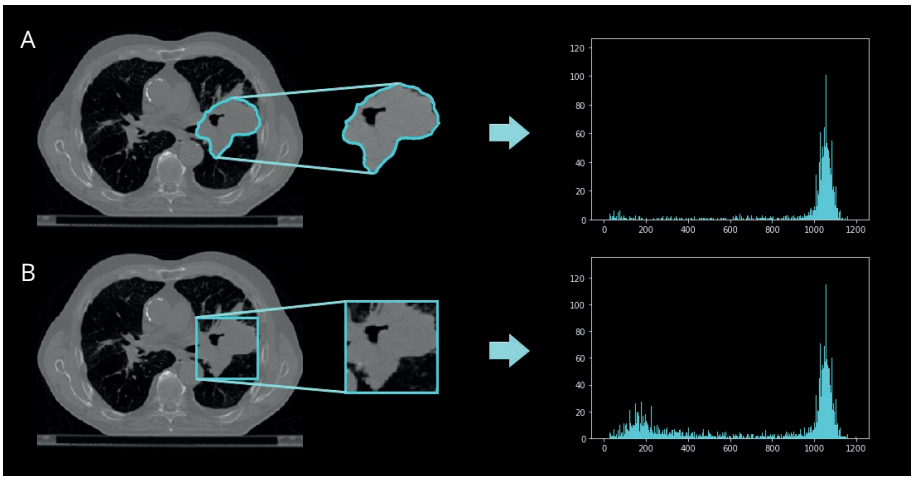


Figure 2. The difference between using (A) a contoured binary mask, and (B) using a bounding box.

Once the ROI is defined, the choice of features to be extracted depend on the information being sought. Shape features such as volume relate only to the definition of the ROI, and if this is manually created, suffer from inter- and intraobserver variability.[47] First-order features give insight into the distribution of pixel intensities, *e.g.* histograms of pixel intensities are quantified by a large number of statistical methods, including variance, skewness, and kurtosis. These features, however, are unable to quantify how pixels are positioned in relation to each other. Second and higher-order features may capture this relationship, with second-order features obtained based on the average relationship between two pixels/voxels, and higher-order features for more than two pixels/voxels. An example of a second-order feature extraction method is the grey level co-occurring matrix (GLCM). GLCMs are co-occurring pixels in each defined direction (Figure 3) and are counted and recorded (Figure 4) into a matrix. Statistical analysis such as contrast, correlation, and homogeneity, as well as tailored formulae can then be applied on the GLCM to extract independent features[48]. Features extracted in this manner are considered "hand crafted" features as they are features that are pre-defined by specially designed formulae.

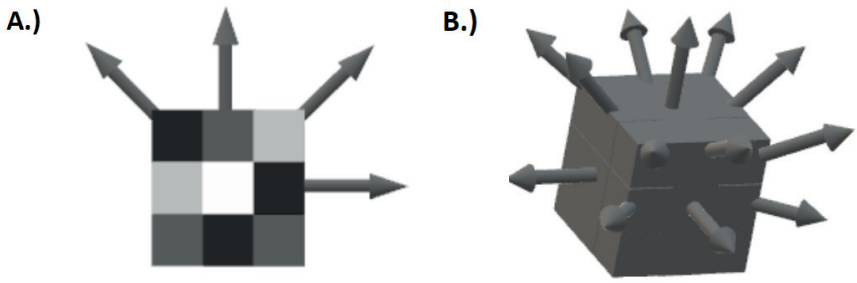


Figure 3. Possible angles for the calculation of co-occurrence matrices in two and three dimensions. (A) Shows the 4 possible directions in 2 dimensions while (B) shows the 13 possible directions in 3 dimensions.

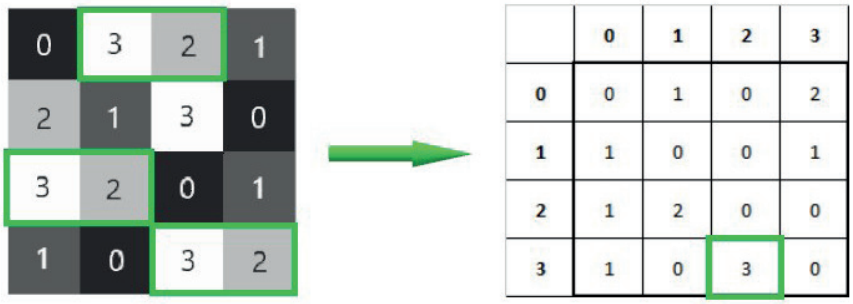


Figure 4. Calculating a GLCM for horizontal co-occurring pixel intensities. In total, 3 co- occurring pixel intensities of 3 and 2 that are next to each other on a horizontal plane can be totalled and tracked in the corresponding matrix. GLCM, grey level co- occurring matrix.

After features have been extracted from all the images in a database, a subset of features needs to be selected that go into the final model. To make a model generalisable, it is important to avoid finding spurious correlations in the data that do not generalise to other similar datasets, an occurrence termed overfitting.[49-51] If a model has learned to recognize noise, outliers, or other kinds of variance, it is unlikely to perform well when presented new data. The larger the number of predictors, the larger the chance to find spurious correlations, a major problem in the realm of machine learning.[52] To detect overfitting, ideally, a model's performance is validated in external datasets with similar population and outcome distributions, but from different centers -- if the model performs significantly better on the training set than on the validation set, overfitting is likely.[53,54] In the absence of an external validation dataset, data can be split into different subsets, and the model trained in one group and validated on the other(s) in a process called cross-validation (Figure 5).[55] During this process, the model hyper-parameters (settings within the model itself, e.g. degree of polynomial fitting) can be further tuned to increase performance in the training and validation sets.[56]

Figure 5. An example of fivefold cross- validation which can be used to evaluate machine learning models. Cross- validation gives the ability to test the result across the entirety of a data set, giving a better estimation of a model's overall performance.

A method to overcome overfitting is to reduce the number of predictors, in this case, imaging features. Feature selection is the process of reducing the number of predictors while retaining the core important information that correlates with outcomes or the underlying biology.[32] Many feature reduction methods exist, but none are known to work well on all kinds of datasets, and they can be combined in many ways.[32] This remains an active field of research.[57] Similar features can also be grouped to achieve dimensionality reduction, and methods such as principal component analysis and independent component analysis are employed to this end.[58]

Once features are selected, the task is to correlate these features - individually or in groups - to diagnostic and prognostic outcomes or to the underlying biology. There are numerous methods to find and test such models, from simple linear regression and curve-fitting to advanced machine learning methods such as decision trees, support vector machines, random forests, boosted trees, or neural networks.[59] Assembling is the combination of models that get trained on random samples of data from the training set called bags and then combined as a whole using a voting system. This is the basis for algorithms such as Random Forests, AdaBoost, and Gradient Boosting.[60] An intuitive explanation is that even though the individual models can show a large amount of variance due to being trained on small subsets of the data, their averaging or voting smooths out the variance while improving the ability to better generalise.[60]

Once a generalisable model has been trained and externally validated, it might be desirable to expand the interoperability of the model to all hardware, acquisition, and reconstruction parameters found in general clinical practice. Instead of relying

on the standardisation of images, the features themselves can be harmonized to a common frame-of-reference using combined batch methods such as ComBat,[44,60,61] originally developed for similar problems encountered in gene sequencing assays.[62]

Deep learning for fully automated workflows

Artificial neural networks (ANNs) are a class of machine learning architecture that are loosely based on how biological brains work.[63] With the exception of unsupervised learning (such as autoencoders), deep learning architectures usually rely on information regarding the outcome in order to craft their features, and unlike in handcrafted radiomics, feature extraction and correlation are intertwined.[64] Also, unlike radiomics, there is generally no need for image segmentation, as the whole image can be presented to a deep learning model, both during training and in clinical routine.

An ANN is able to use a collection of neurons and weights, one for each of the inputs preceding the neuron.[65] These weights get continuously updated, or corrected, in steps called epochs that work together to create a very complex function able to make predictions. The weights are inputs for each neuron and are multiplied and averaged, resulting in a transfer function, which is converted to an output via a function called an activation function.[66] These activation functions are often a sigmoidal function such as a hyperbolic tangent or sigmoid, or a function called a rectified linear unit that can be represented as the maximum of the product of the coefficient and zero or one. A representation of a single neuron, including the activation function, can be seen in Figure 6.[67] Multiple neurons can then be stacked to create a single layer referred to as a "hidden layer" and hidden layers (were inputs and outputs all connect) can be stacked to create larger networks, see Figure 7.[65] The term deep learning is used to describe a neural network that has many layers, which is considered deep. For a binary classifier or regression, the final layer should contain only a single neuron and use a sigmoid activation function to make a prediction with a binary outcome (zero or one). If the problem is categorical, the network's final layer should contain the same number of neurons as there are categories to be classified and the final activation will be a "softmax" function, which is the average of the exponentials of the inputs,[68] yielding the probabilities of each category. Deep learning for image vision employs convolutional neural networks (CNNs) which are a type of ANN that have an automated feature extractor designed specifically for images.[69] CNNs employ a filtering technique, which convolves the image with a kernel (sliding window), creating a new pixel/voxel value (and hence new image) by sliding a matrix of numbers over the image, see Figure 8. It is possible to make a variety of different filters using these types of convolutions, such as blurring, sharpening, edge detection, and gradient detection,[69,70] and CNNs are able to learn filters that are best suited to extracting features needed for making predictions.

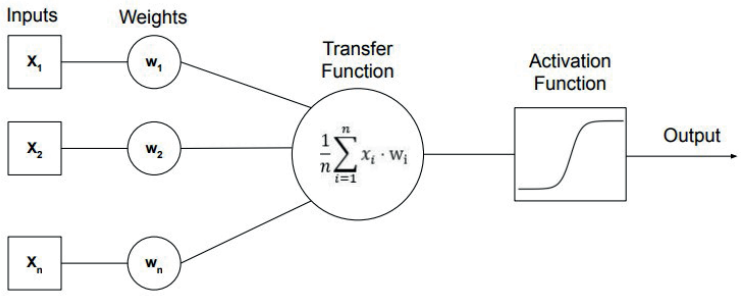


Figure 6. The architecture of a single neuron with a transfer function and a sigmoid activation function visualised.

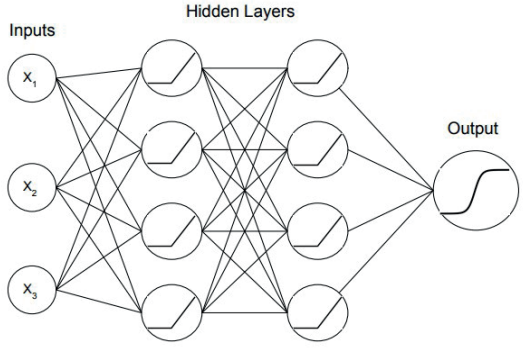


Figure 7. A three-layer neural network that is a binary classifier with three inputs. Nodes with Xn refer to inputs while other nodes refer to activation functions. The connecting lines between the nodes represent weights.

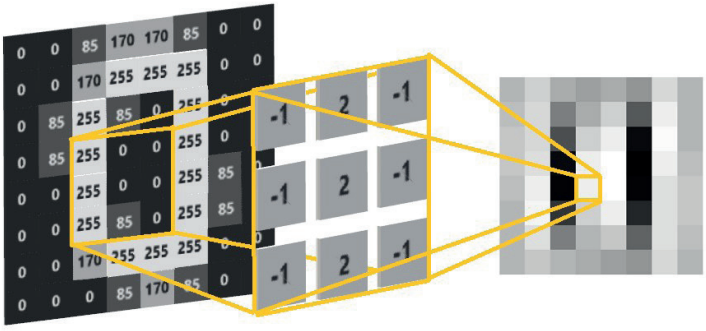


Figure 8. A filter that is able to filter out vertical lines. The yellow lines represent the kernel or sliding window, while the image on the right is the result of performing convolutions across the entirety of the original image.

ANNs do have some drawbacks compared to using hand crafted features alongside other machine learning techniques. The main drawback is the intrinsic need for much larger datasets to train the models, since feature creation is contingent on the training data, as opposed to handcrafted radiomics. Another drawback to using ANNs is interpretability. ANNs build ultra-complex functions that can be extremely difficult for practitioners to make sense of. Although CNNs have performed very well in image recognition, they have been less successful learning texture features, since texture information inherently has a higher dimensionality compared to other types of datasets, making them more difficult for neural networks to master.[69,71] According to Basu et al (2018), a redesign of neural network architectures is required to extract features in a similar manner as GLCM and other features based on spatial correlation.

Currently, the main application of deep learning in the radiomics workflow still lies in the automated detection and localization of organs and lesions, removing the major burden in dataset curation. While there is no algorithm that can solve every problem, deep learning still has its place and is able to work as additional methods for delineation and feature extraction that compliments handcrafted radiomics. There is active research in combining both deep learning features and radiomics features that shows improved results.[72-74]

Potential Clinical Applications

Radiomics in Oncology

Radiomics has been widely studied for application in diagnosis and treatment prognosis/selection in oncology, primarily due to the existence of large imaging datasets used for staging, often containing delineations of tumours and organs at risk necessary for radiation treatment planning. These datasets can be used to train diagnostic and prognostic models for a variety of cancer types and sites. Using clinical reports, pathology/histology, and genetic information along with radiomics analysis can give a global outlook on the biology of the disease.[48] In this section, an overview of notable studies published in this area will be discussed.

Lung

Lung cancer is by far the leading cause of cancer-related deaths among both males and females worldwide.[75] Recent studies have shown that radiomics can determine the risk of lung cancer from screening scans.[76-78] Radiomic features found to have a strong association to decode tumour heterogeneity for risk stratification,[79,80] concluding that patients with heterogeneous tumours tend to have a worse prognosis. In addition to that, Yoon et al. were able to show the association of radiomic analysis with gene expression.[80] Radiomic features were

also found to correlate with TNM staging for lung and head-and-neck cancer.[31,81] Later studies further validated the strong predictive power of radiomics for distant metastasis.[82-84]

Radiomics may also play a role in lung cancer treatment planning by evaluating tumour response to a specific treatment. Several studies focused on analysing the tumour response to radiation therapy.[85,86] For instance, Mattonen et al. developed a radiomics signature for treatment response to stereotactic ablative radiation therapy that was able to predict lung cancer recurrence post-therapy,[85] while Fave et al. used multiple time point information referred to as delta-radiomic analysis to evaluate the change of radiomic features as a predictor for tumour response to radiation therapy.[86] The results suggest that delta radiomic features are in fact a good indicator of treatment response. Another interesting study by Mattonen et al. found that radiomic analysis can identify features associated with local recurrence of lung cancer after radiation therapy,[87] while physicians usually have great difficulty to distinguish local recurrence from radiation-induced sequelae.

Besides the traditional handcrafted feature extraction approach followed in the radiomics pipeline, deep learning radiomics is also gaining popularity among researchers. A deep learning-based approach followed by Shen et al. yielded more accurate malignancy prediction of nodules compared to previous methods.[88] Pham et al. used a two-step deep learning approach for evaluating lymph node metastases with accurate cancer detection.[89] Instead of using data from a single time point, deep recurrent convolutional network architectures can be used to analyse data from multiple time points to monitor treatment response.[90]

Brain

Brain tumours are usually graded based on clinical or pathological analysis to define their malignancy. Radiomics may be able to non-invasively perform grade assessment, as reported by Coroller et al. in meningioma patients, suggesting a strong correlation between certain imaging features and histopathologic grade.[91] Zhang et al. were able to classify between low-grade gliomas and high-grade gliomas with high accuracy.[92] Chen et al. investigated the prediction of brain metastases in T1 lung adenocarcinoma patients and found that the predictive performance for the radiomics model was significantly better compared to clinical models and could potentially be used for brain metastases screening.[93] Fetit et al. performed radiomic analysis for the classification of brain tumours in childhood suggesting that radiomics can aid in the classification of tumour subtype.[94] However, the scalability of the techniques used in these studies needs to be assessed further by extensions to multicentric cohorts using different acquisition protocols and vendors.

Radiation therapy can lead to necrosis, which is difficult to distinguish from tumour recurrence on imaging. Larroza et al. were able to develop a high classification accuracy model to distinguish between brain metastasis and radiation necrosis using radiomic analysis.[95] Some radiomic studies successfully investigated the treatment response in recurrent glioblastoma patients with a radiomics approach.[96-98] An iterative study by radiomic researchers found strong evidence of radiomic features in predicting survival and treatment response of patients with glioblastoma using pre-treatment imaging data.[99-101]

Deep learning has also made some other interesting contributions in this area. Chang et al. used residual deep convolutional network for predicting the genotype in grade II-IV glioma with high accuracy.[102] Deep learning can also be used complementary to traditional hand crafted radiomics studies. For example, studies[72,73] focused on using deep networks for segmentation, followed by radiomics analysis for survival prediction.

Breast

Among women, breast cancer is the second leading cause of death for cancer worldwide.[75] However, earlier diagnosis can lead to a better prognosis. Radiomics in the field of breast cancer has been applied to several imaging modalities including (PET)-MRI, (contrast-enhanced) mammography, ultrasound, and digital breast tomosynthesis focusing on tumour classification, molecular subtypes, tumour response prediction to neoadjuvant systemic therapy (NST), lymph node metastasis, overall survival, and recurrence risks. For example, a large number of radiomics studies have been used for the prediction of malignant breast cancers.[103-106] Besides the prediction of tumour malignancy, several radiomics studies examined the prediction of breast cancer molecular subtypes with the aim of leaving out liquid biopsies in the future.[107-110] Lymph node metastasis identification is an important prognostic factor and often determines treatment. In all clinically node negative patients, a sentinel lymph node procedure is the basis of the axillary treatment.[111] Dong et al. was able to provide an alternative to this invasive approach by successfully applying radiomics for the prediction of lymph node metastasis in the sentinel lymph node using imaging data.[102]

In addition to the prediction of breast tumour malignancy, tumour molecular subtypes and sentinel lymph node metastasis identification, radiomics studies have also made some significant contributions to treatment planning. Chan et al. investigated the power of radiomics to discriminate between patients with low and high treatment failure risk on pre-treatment imaging data.[112] There are multiple studies that predict tumour response to neoadjuvant systemic therapy using radiomic analysis. For instance, Braman et al. found a combination of intratumoural and peritumoural radiomics features as a robust and strong indicator for pathologic

complete tumour response using pre-treatment imaging data.[113] Two other studies[114,115] found similar evidence on serial imaging data containing follow-up scans. The use of multiparametric MRI for the prediction of tumour response to NST showed promising results.[116,117]

Deep learning approaches have also been adopted in breast cancer research. The study of Huynh et al. investigated tumour classification capacity of deep features extracted from convolutional networks trained on a different dataset to analytically extracted features.[118] The results suggested a higher performance of deep features. Similarly, another study,[119] used deep learning for risk assessment and found higher performance compared to conventional texture analysis.

Other sites and diseases

While cancers of the lung, brain, and breast have received wide attention from the radiomics research community, any site is open to QIA research. Diagnostic and prognostic radiomics research is ongoing for cancers of the head-and-neck,[120] ovaries,[38] prostate,[121] kidney,[122] liver,[123] colon and rectum,[124] and many other sites. The main requirements for a radiomics study are the presence of a radiologic phenotype which allows for the clustering of patients based on differences within that phenotype or some correlation to the underlying biology, and the availability of imaging and clinical data. While not nearly as prevalent,[125] this has meant that non-oncological diseases which require medical imaging as part of the standard of care have also been the subject of radiomics analysis, such as in the fields of neurology,[35] ophthalmology,[126] and dentistry.[127]

Limitations of radiomics and future directions towards precision medicine

While radiomics facilitates new possibilities in the field of personalised medicine, some challenges remain. One of the primary obstacles is the lack of big and standardised clinical data. Although large amounts of medical imaging data are stored, these data are dispersed across different centers and acquired using different protocols. Access for research purposes is highly restricted by law and ethics. An exhaustive data curation and harmonization process is still necessary to make it usable for research. Radiomics will potentially enable imaging-based clinical decision support systems, however, the current black box approach, particularly in deep learning, makes it less acceptable for clinical application. In certain cases, hand crafted radiomic features have already been correlated with biological processes,[128-130] but it is essential to work further in the direction of interpretable artificial intelligence (AI) to make it more accessible for clinical implementation [33].

In recent years, various countries have already adopted many measures to control variability in clinical trial protocols, data acquisition, and analysis.[131,132] For example, across Europe consistent protocol guidance was adopted with the help of European Association of Nuclear Medicine.[133] The Quantitative Imaging Biomarker Alliance initiative also aims to achieve the same task in a much broader level.[134,135] On the other hand, algorithmically, developments in deep learning allow for automated quality check, clustering of data, and automated detection and contouring of organs and lesions, vastly improving data curation times. Generative adversarial networks open up the possibility of generating synthetic data[136] or domain adaptive algorithms[137,138] might be able to deal with the shortage of standardized data. Techniques like distributed learning provide the ability to train machine learning models using distributed data without the data ever leaving their original locations. Distributed learning has already been applied across several medical institutions to build predictive and segmentation models.[139-142] Furthermore, this approach can be coupled with other technologies such as blockchain to trace back data provenance and monitor the use of the final models.[143] Various techniques to visualize deep features have already been put forward by researchers to generate an intuitive understanding. A completely new research area of Artificial Intelligence called explainable AI aims to track the decisions made by the intelligent algorithms so that it can be better understood by humans. Companies like Google, IBM, Microsoft and Facebook are at the forefront in this research. This will not only helps to build trust of AI systems among medical professionals but also unlocks new possibilities in understanding a disease.[144,145]

The implementation of precision medicine itself has its own limitations and has drawn criticism due to the lack of a "transformation in therapeutic medicine" in the last two decades.[146] So far life expectancies or other public health measures have not shown any dramatic improvements, regardless of the vast amounts of precision medicine research being conducted. Contentious points remain such as excessive costs (*e.g.* gene therapy), although new developments such as radiomics promise to reduce costs in the long run. Furthermore, the diagnostic and prognostic power of complex "omics-driven" models is still to be determined in specific populations, and evidence needs to be produced that such methods improve health outcomes.[147] Precision medicine is likely to mature and translate to clinical workflows over the next decade and will change the way health services are delivered and evaluated. Healthcare systems will need to adjust their methods and processes to accommodate for these changes.

Conclusion

Radiomics, whether handcrafted or deep, is an emerging field that translates medical images into quantitative data to give biological information and enable phenotypic profiling for diagnosis, theragnosis, decision support, and monitoring. Radiomics, in essence, allows personalised care by identifying features or signatures correlated with a disease or a treatment response with high precision and in a non-invasive way. Recent developments in genomics and deep learning have pushed radiomics researchers to focus more on extracting deep features and explore new possibilities in artificial intelligence modelling. In the future, radiomics will be a valued addition to precision medicine workflows by facilitating earlier and more accurate diagnosis, providing prognostic information, aiding in treatment choice, monitoring disease and treatment non-invasively, and enabling routine dynamic treatment based on individual responses. But the road to this vision is long, and many technical, regulatory, and ethical problems still need to be solved.

References

1. Scatliff JH, Morris PJ. From Roentgen to magnetic resonance imaging: the history of medical imaging. N C Med J. 2014 Mar;75(2):111–3.
2. Giakos GC, Pastorino M, Russo F, Chowdhury S, Shah N, Davros W. Noninvasive imaging for the new century [Internet]. Vol. 2, IEEE Instrumentation & Measurement Magazine. 1999. p. 32–5, 49. Available from: http://dx.doi.org/10.1109/5289.765967
3. Prince J, Links J. Medical Imaging: Signals and Systems (Prince, J.L. and Links, J.M.; 2006) [Book Review] [Internet]. Vol. 25, IEEE Signal Processing Magazine. 2008. p. 152–3. Available from: http://dx.doi.org/10.1109/msp.2008.4408454
4. Kesner A, Laforest R, Otazo R, Jennifer K, Pan T. Medical imaging data in the digital innovation age. Med Phys. 2018 Apr;45(4):e40–52.
5. Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev. 1956 Mar;63(2):81–97.
6. Wang Y-XJ, Ng CK. The impact of quantitative imaging in medicine and surgery: Charting our course for the future. Quant Imaging Med Surg. 2011 Dec;1(1):1–3.
7. Lodwick GS, Haun CL, Smith WE, Keller RF, Robertson ED. Computer Diagnosis of Primary Bone Tumors [Internet]. Vol. 80, Radiology. 1963. p. 273–5. Available from: http://dx.doi.org/10.1148/80.2.273
8. Meyers PH, Nice CM. Automated Computer Analysis of Radiographic Images [Internet]. Vol. 8, Archives of Environmental Health: An International Journal. 1964. p. 774–5. Available from: http://dx.doi.org/10.1080/00039896.1964.10663755
9. Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W. Detection of Radiographic Abnormalities in Mammograms by Means of Optical Scanning and Computer Analysis [Internet]. Vol. 89, Radiology. 1967. p. 211–5. Available from: http://dx.doi.org/10.1148/89.2.211
10. Summers RM. Road maps for advancement of radiologic computer-aided detection in the 21st century. Radiology. 2003 Oct;229(1):11–3.
11. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph. 2007 Jun;31(4-5):198–211.
12. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228–47.
13. Zheng B. Identifying and testing new quantitative image, an analysis based clinical markers to predict breast cancer risk and prognosis [Internet]. Vol. 05, OMICS Journal of Radiology. 2016. Available from: http://dx.doi.org/10.4172/2167-7964.c1.009
14. Kobayashi T, Xu XW, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology. 1996 Jun;199(3):843–8.
15. Halalli B, Makandar A. Computer Aided Diagnosis - Medical Image Analysis Techniques [Internet]. Breast Imaging. 2018. Available from: http://dx.doi.org/10.5772/intechopen.69792
16. The Robust Computer Aided Diagnostic System for Lung Nodule Diagnosis [Internet]. Vol. 8, International Journal of Recent Technology and Engineering. 2019. p. 5670–5. Available from: http://dx.doi.org/10.35940/ijrte.d8169.118419

17. Ziyad SR, Radha V, Vayyapuri T. Overview of Computer Aided Detection and Computer Aided Diagnosis Systems for Lung Nodule Detection in Computed Tomography [Internet]. Vol. 16, Current Medical Imaging Formerly Current Medical Imaging Reviews. 2020. p. 16–26. Available from: http://dx.doi.org/10.2174/1573405615666190206153321
18. Rizzi M, D'Aloia M, Castagnolo B. Computer aided detection of microcalcifications in digital mammograms adopting a wavelet decomposition [Internet]. Vol. 16, Integrated Computer-Aided Engineering. 2009. p. 91–103. Available from: http://dx.doi.org/10.3233/ica-2009-0306
19. Gibbs P, Turnbull LW. Textural analysis of contrast-enhanced MR images of the breast. Magn Reson Med. 2003 Jul;50(1):92–8.
20. Murakami R, Kumita S, Tani H, Yoshida T, Sugizaki K, Kuwako T, et al. Detection of breast cancer with a computer-aided detection applied to full-field digital mammography. J Digit Imaging. 2013 Aug;26(4):768–73.
21. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. Theranostics. 2019 Feb 12;9(5):1303–22.
22. Hunter LA, Krafft S, Stingo F, Choi H, Martel MK, Kry SF, et al. High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. Med Phys. 2013;40(12):121916.
23. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings. PLoS One. 2016 Dec 29;11(12):e0166550.
24. He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule [Internet]. Vol. 6, Scientific Reports. 2016. Available from: http://dx.doi.org/10.1038/srep34921
25. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography. 2016 Dec;2(4):361–5.
26. El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit. 2009 Jun 1;42(6):1162–71.
27. Ulaner GA. Measuring Treatment Response on FDG PET/CT [Internet]. Fundamentals of Oncologic PET/CT. 2019. p. 225–9. Available from: http://dx.doi.org/10.1016/b978-0-323-56869-2.00022-3
28. Xu Q-G, Xian J-F. Role of quantitative magnetic resonance imaging parameters in the evaluation of treatment response in malignant tumors. Chin Med J . 2015 Apr 20;128(8):1128–33.
29. Degnan AJ, Chung CY, Shah AJ. Quantitative diffusion-weighted magnetic resonance imaging assessment of chemotherapy treatment response of pediatric osteosarcoma and Ewing sarcoma malignant bone tumors [Internet]. Vol. 47, Clinical Imaging. 2018. p. 9–13. Available from: http://dx.doi.org/10.1016/j.clinimag.2017.08.003
30. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012 Mar;48(4):441–6.
31. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014 Jun 3;5:4006.

32. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine [Internet]. Vol. 14, Nature Reviews Clinical Oncology. 2017. p. 749–62. Available from: http://dx.doi.org/10.1038/nrclinonc.2017.141
33. Sanduleanu S, Woodruff HC, de Jong EEC. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother Oncol [Internet]. 2018; Available from: https://www.sciencedirect.com/science/article/pii/S0167814018301798
34. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys. 2018 Jul;45(7):3449–59.
35. Ibrahim A, Vallières M, Woodruff H, Primakov S, Beheshti M, Keek S, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine [Internet]. Vol. 49, Seminars in Nuclear Medicine. 2019. p. 438–49. Available from: http://dx.doi.org/10.1053/j.semnuclmed.2019.06.005
36. Refaee T, Wu G, Ibrahim A, Halilaj I, Leijenaar RTH, Rogers W, et al. The Emerging Role of Radiomics in COPD and Lung Cancer. Respiration [Internet]. 2020 Jan 28; Available from: http://dx.doi.org/10.1159/000505429
37. Collins FS. Medical and Societal Consequences of the Human Genome Project [Internet]. Vol. 341, New England Journal of Medicine. 1999. p. 28–37. Available from: http://dx.doi.org/10.1056/nejm199907013410106
38. Nougaret S, Tardieu M, Vargas HA, Reinhold C, Vande Perre S, Bonanno N, et al. Ovarian cancer: An update on imaging in the era of radiomics. Diagn Interv Imaging. 2019 Oct;100(10):647–55.
39. Giger ML, Doi K, MacMahon H, Dwyer SJ III, Schneider RH. Computerized Detection Of Lung Nodules In Digital Chest Radiographs [Internet]. Medical Imaging. 1987. Available from: http://dx.doi.org/10.1117/12.967022
40. Giger ML, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. Med Phys. 1988 Mar;15(2):158–66.
41. Larue RTHM, Van De Voorde L, van Timmeren JE, Leijenaar RTH, Berbée M, Sosef MN, et al. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. Radiother Oncol. 2017 Oct;125(1):147-53.
42. Bagher-Ebadian H, Siddiqui F, Liu C, Movsas B, Chetty IJ. On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. Med Phys. 2017 May;44(5):1755–70.
43. Haga A, Takahashi W, Aoki S, Nawa K, Yamashita H, Abe O, et al. Standardization of imaging features for radiomics analysis. J Med Invest. 2019;66(1.2):35–7.
44. Kalpathy-Cramer. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. TOMOGRAPH. 2016 Dec;2(4).
45. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One. 2014 Jul 15;9(7):e102107.
46. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpt WJC, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013 Oct;52(7):1391–7.
47. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb;278(2):563–77.

48. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018 Nov 14;2(1):36.

49. Vial A, Stirling D, Field M, Ros M, Ritz C, Carolan M, et al. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review [Internet]. Vol. 7, Translational Cancer Research. 2018. p. 803–16. Available from: http://dx.doi.org/10.21037/tcr.2018.05.02

50. Fan J, Shao Q-M, Zhou W-X. ARE DISCOVERIES SPURIOUS? DISTRIBUTIONS OF MAXIMUM SPURIOUS CORRELATIONS AND THEIR APPLICATIONS. Ann Stat. 2018 Jun;46(3):989–1017.

51. Lever J, Krzywinski M, Altman N. Model selection and overfitting [Internet]. Vol. 13, Nature Methods. 2016. p. 703–4. Available from: http://dx.doi.org/10.1038/nmeth.3968

52. Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts HJWL. Data Analysis Strategies in Medical Imaging. Clin Cancer Res. 2018 Aug 1;24(15):3492–9.

53. Chatterjee A, Vallières M, Dohan A, Levesque IR, Ueno Y, Bist V, et al. An Empirical Approach for Avoiding False Discoveries When Applying High-Dimensional Radiomics to Small Datasets. IEEE Transactions on Radiation and Plasma Medical Sciences. 2019 Mar;3(2):201–9.

54. Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation [Internet]. Vol. 48, Pattern Recognition. 2015. p. 2839–46. Available from: http://dx.doi.org/10.1016/j.patcog.2015.03.009

55. Duarte E, Wainer J. Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters [Internet]. Vol. 88, Pattern Recognition Letters. 2017. p. 6–11. Available from: http://dx.doi.org/10.1016/j.patrec.2017.01.007

56. Solorio-Fernández S, Ariel Carrasco-Ochoa J, Martínez-Trinidad JF. A review of unsupervised feature selection methods [Internet]. Artificial Intelligence Review. 2019. Available from: http://dx.doi.org/10.1007/s10462-019-09682-y

57. Zhang D, Zou L, Zhou X, He F. Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer [Internet]. Vol. 6, IEEE Access. 2018. p. 28936–44. Available from: http://dx.doi.org/10.1109/access.2018.2837654

58. Choudhary R, Gianey HK. Comprehensive Review On Supervised Machine Learning Algorithms [Internet]. 2017 International Conference on Machine Learning and Data Science (MLDS). 2017. Available from: http://dx.doi.org/10.1109/mlds.2017.11

59. Ren Y, Zhang L, Suganthan PN. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article] [Internet]. Vol. 11, IEEE Computational Intelligence Magazine. 2016. p. 41–53. Available from: http://dx.doi.org/10.1109/mci.2015.2471235

60. Lovinfosse P, Visvikis D, Hustinx R, Hatt M. FDG PET radiomics: a review of the methodological aspects [Internet]. Vol. 6, Clinical and Translational Imaging. 2018. p. 379–91. Available from: http://dx.doi.org/10.1007/s40336-018-0292-9

61. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118–27.

62. Lucia F, Visvikis D, Vallières M, Desseroit M-C, Miranda O, Robin P, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2019 Apr;46(4):864–77.

63. Kriegeskorte N. Deep neural networks: a new framework for modelling biological vision and brain information processing [Internet]. Available from: http://dx.doi.org/10.1101/029876

64. Bengio Y, Delalleau O, Le Roux N. The Curse of Highly Variable Functions for Local Kernel Machines. Adv Neural Inf Process Syst. 2005;18:107–14.

65. Hinton GE. Learning multiple layers of representation. Trends Cogn Sci. 2007 Oct;11(10):428–34.

66. LeCun Y. Deep learning & convolutional networks [Internet]. 2015 IEEE Hot Chips 27 Symposium (HCS). 2015. Available from: http://dx.doi.org/10.1109/hotchips.2015.7477328

67. LeCun Y, Bengio Y, Hinton G. Deep learning [Internet]. Vol. 521, Nature. 2015. p. 436–44. Available from: http://dx.doi.org/10.1038/nature14539

68. Aggarwal CC. Neural Networks and Deep Learning: A Textbook. Springer; 2018. 497 p.

69. Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: A review. Neurocomputing. 2016 Apr 26;187:27–48.

70. Tyagi V. Introduction to Digital Image Processing [Internet]. Understanding Digital Image Processing. 2018. p. 1–12. Available from: http://dx.doi.org/10.1201/9781315123905-1

71. Basu S, Mukhopadhyay S, Karki M, DiBiano R, Ganguly S, Nemani R, et al. Deep neural networks for texture classification—A theoretical analysis [Internet]. Vol. 97, Neural Networks. 2018. p. 173–82. Available from: http://dx.doi.org/10.1016/j.neunet.2017.10.001

72. Yogananda CGB, Nalawade SS, Murugesan GK, Wagner B, Pinho MC, Fei B, et al. Fully Automated Brain Tumor Segmentation and Survival Prediction of Gliomas using Deep Learning and MRI [Internet]. Available from: http://dx.doi.org/10.1101/760157

73. Sun L, Zhang S, Chen H, Luo L. Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning. Front Neurosci. 2019 Aug 16;13:810.

74. A. Jochems, R. T. H. Leijenaar, M. Bogowicz, F. J. P. Hoebers, F. Wesseling, S. H. Huang, B. Chan, J. N. Waldron, B. O'Sullivan, D. Rietveld, C. R. Leemans, O. Riesterer, S. Tanadini-Lang, M. Guckenberger, K. Ikenberg, P. Lambin. Combining deep learning and radiomics to predict HPV status in oropharyngeal squamous cell carcinoma. Radiotherapy & Oncology. 2018 Apr;127:S504–5.

75. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394–424.

76. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, et al. Predicting Malignant Nodules from Screening CT Scans. J Thorac Oncol. 2016 Dec;11(12):2120–8.

77. Kumar D, Chung AG, Shaifee MJ, Khalvati F, Haider MA, Wong A. Discovery Radiomics for Pathologically-Proven Computed Tomography Lung Cancer Prediction [Internet]. Lecture Notes in Computer Science. 2017. p. 54–62. Available from: http://dx.doi.org/10.1007/978-3-319-59876-5_7

78. Liu Y, Balagurunathan Y, Atwater T, Antic S, Li Q, Walker RC, et al. Radiological Image Traits Predictive of Cancer Status in Pulmonary Nodules. Clin Cancer Res. 2017 Mar 15;23(6):1442–9.

79. Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S, Miles K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol. 2012 Apr;22(4):796–802.

80. Yoon HJ, Sohn I, Cho JH, Lee HY, Kim J-H, Choi Y-L, et al. Decoding Tumor Phenotypes for ALK, ROS1, and RET Fusions in Lung Adenocarcinoma Using a Radiomics Approach. Medicine . 2015 Oct;94(41):e1753.

81. Yoon HJ, Sohn I, Cho JH, Lee HY, Kim J-H, Choi Y-L, et al. Decoding Tumor Phenotypes for ALK, ROS1, and RET Fusions in Lung Adenocarcinoma Using a Radiomics Approach. Medicine . 2015 Oct;94(41):e1753.

82. Parmar C, Leijenaar RTH, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Sci Rep. 2015 Jun 5;5:11044.

83. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol. 2015 Mar;114(3):345–50.

84. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW Jr, et al. Early-Stage Non-Small Cell Lung Cancer: Quantitative Imaging Characteristics of (18)F Fluorodeoxyglucose PET/CT Allow Prediction of Distant Metastasis. Radiology. 2016 Oct;281(1):270–8.

85. Zhou H, Dong D, Chen B, Fang M, Cheng Y, Gan Y, et al. Diagnosis of Distant Metastasis of Lung Cancer: Based on Clinical and Radiomic Features. Transl Oncol. 2018 Feb;11(1):31–6.

86. Mattonen SA, Tetar S, Palma DA, Senan S, Ward AD. Automated Texture Analysis for Prediction of Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer [Internet]. Vol. 93, International Journal of Radiation Oncology*Biology*Physics. 2015. p. S5–6. Available from: http://dx.doi.org/10.1016/j.ijrobp.2015.07.019

87. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Sci Rep. 2017 Apr 3;7(1):588.

88. Mattonen SA, Palma DA, Johnson C, Louie AV, Landis M, Rodrigues G, et al. Detection of Local Cancer Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer: Physician Performance Versus Radiomic Assessment. Int J Radiat Oncol Biol Phys. 2016 Apr 1;94(5):1121–8.

89. Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, et al. Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification [Internet]. Vol. 61, Pattern Recognition. 2017. p. 663–73. Available from: http://dx.doi.org/10.1016/j.patcog.2016.05.029

90. Pham HHN, Futakuchi M, Bychkov A, Furukawa T, Kuroda K, Fukuoka J. Detection of lung cancer lymph node metastases from whole-slide histopathological images using a two-step deep learning approach. Am J Pathol [Internet]. 2019 Sep 18; Available from: http://dx.doi.org/10.1016/j.ajpath.2019.08.014

91. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. Clin Cancer Res. 2019 Jun 1;25(11):3266–75.

92. Coroller TP, Bi WL, Huynh E, Abedalthagafi M, Aizer AA, Greenwald NF, et al. Radiographic prediction of meningioma grade by semantic and radiomic features. PLoS One. 2017 Nov 16;12(11):e0187908.

93. Zhang X, Yan L-F, Hu Y-C, Li G, Yang Y, Han Y, et al. Optimizing a machine learning based glioma grading system using multi-parametric MRI histogram and texture features. Oncotarget. 2017 Jul 18;8(29):47816–30.

94. Chen A, Lu L, Pu X, Yu T, Yang H, Schwartz LH, et al. CT-Based Radiomics Model for Predicting Brain Metastasis in Category T1 Lung Adenocarcinoma. AJR Am J Roentgenol. 2019 Apr 1;1–6.

95. Fetit AE, Novak J, Peet AC, Arvanitits TN. Three-dimensional textural features of conventional MRI improve diagnostic classification of childhood brain tumours. NMR Biomed. 2015 Sep;28(9):1174–84.

96. Larroza A, Moratal D, Paredes-Sánchez A, Soria-Olivas E, Chust ML, Arribas LA, et al. Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in MRI. J Magn Reson Imaging. 2015 Nov;42(5):1362–8.

97. Kickingereder P, Götz M, Muschelli J, Wick A, Neuberger U, Shinohara RT, et al. Large-scale Radiomic Profiling of Recurrent Glioblastoma Identifies an Imaging Predictor for Stratifying Anti-Angiogenic Treatment Response. Clin Cancer Res. 2016 Dec 1;22(23):5765–71.

98. Chang K, Zhang B, Guo X, Zong M, Rahman R, Sanchez D, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. Neuro Oncol. 2016 Dec;18(12):1680–7.

99. Grossmann P, Narayan V, Chang K, Rahman R, Abrey L, Reardon DA, et al. Quantitative imaging biomarkers for risk stratification of patients with recurrent glioblastoma treated with bevacizumab [Internet]. Vol. 19, Neuro-Oncology. 2017. p. 1688–97. Available from: http://dx.doi.org/10.1093/neuonc/nox092

100. Pérez-Beteta J, Molina D, Martínez-González A, Arregui E, Asenjo B, Iglesias L, et al. P09.43 Novel geometrical imaging biomarkers predict survival and allow for patient selection for surgery in glioblastoma patients [Internet]. Vol. 19, Neuro-Oncology. 2017. p. iii80–iii80. Available from: http://dx.doi.org/10.1093/neuonc/nox036.299

101. Pérez-Beteta J, Molina-García D, Ortiz-Alhambra JA, Fernández-Romero A, Luque B, Arregui E, et al. Tumor Surface Regularity at MR Imaging Predicts Survival and Response to Surgery in Patients with Glioblastoma. Radiology. 2018 Jul;288(1):218–25.

102. Pérez-Beteta J, Molina-García D, Martínez-González A, Henares-Molina A, Amo-Salas M, Luque B, et al. Correction to: Morphological MRI-based features provide pretreatment survival prediction in glioblastoma [Internet]. Vol. 29, European Radiology. 2019. p. 2729–2729. Available from: http://dx.doi.org/10.1007/s00330-018-5870-8

103. Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, et al. Residual Convolutional Neural Network for the Determination of Status in Low- and High-Grade Gliomas from MR Imaging. Clin Cancer Res. 2018 Mar 1;24(5):1073–81.

104. Bickelhaupt S, Paech D, Kickingereder P, Steudle F, Lederer W, Daniel H, et al. Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography [Internet]. Vol. 46, Journal of Magnetic Resonance Imaging. 2017. p. 604–16. Available from: http://dx.doi.org/10.1002/jmri.25606

105. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI [Internet]. Vol. 3, npj Breast Cancer. 2017. Available from: http://dx.doi.org/10.1038/s41523-017-0045-3

106. Bickelhaupt S, Jaeger PF, Laun FB, Lederer W, Daniel H, Kuder TA, et al. Radiomics Based on Adapted Diffusion Kurtosis Imaging Helps to Clarify Most Mammographic Findings Suspicious for Cancer. Radiology. 2018 Jun;287(3):761–70.

107. Whitney HM, Taylor NS, Drukker K, Edwards AV, Papaioannou J, Schacht D, et al. Additive Benefit of Radiomics Over Size Alone in the Distinction Between Benign Lesions and Luminal A Cancers on a Large Clinical Breast MRI Dataset. Acad Radiol. 2019 Feb;26(2):202–9.

108. Guo W, Li H, Zhu Y, Lan L, Yang S, Drukker K, et al. Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data. J Med Imaging (Bellingham). 2015 Oct;2(4):041007.

109. Li H, Zhu Y, Burnside ES, Huang E, Drukker K, Hoadley KA, et al. Quantitative MRI radiomics in

the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ Breast Cancer [Internet]. 2016 May 11;2. Available from: http://dx.doi.org/10.1038/npjbcancer.2016.12

110. Fan M, Li H, Wang S, Zheng B, Zhang J, Li L. Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer [Internet]. Vol. 12, PLOS ONE. 2017. p. e0171683. Available from: http://dx.doi.org/10.1371/journal.pone.0171683
111. Ma W, Zhao Y, Ji Y, Guo X, Jian X, Liu P, et al. Breast Cancer Molecular Subtype Prediction by Mammographic Radiomic Features. Acad Radiol. 2019 Feb;26(2):196–201.
112. Dong Y, Feng Q, Yang W, Lu Z, Deng C, Zhang L, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. Eur Radiol. 2018 Feb;28(2):582–91.
113. Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, et al. Residual Convolutional Neural Network for the Determination of Status in Low- and High-Grade Gliomas from MR Imaging. Clin Cancer Res. 2018 Mar 1;24(5):1073–81.
114. Chan HM, van der Velden BHM, Loo CE, Gilhuijs KGA. Eigentumors for prediction of treatment failure in patients with early-stage breast cancer using dynamic contrast-enhanced MRI: a feasibility study [Internet]. Vol. 62, Physics in Medicine & Biology. 2017. p. 6467–85. Available from: http://dx.doi.org/10.1088/1361-6560/aa7dc5
115. Braman NM, Etesami M, Prasanna P, Dubchuk C, Gilmore H, Tiwari P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. Breast Cancer Res. 2017 May 18;19(1):57.
116. Henderson S, Purdie C, Michie C, Evans A, Lerski R, Johnston M, et al. Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer [Internet]. Vol. 27, European Radiology. 2017. p. 4602–11. Available from: http://dx.doi.org/10.1007/s00330-017-4850-8
117. Parikh J, Selmi M, Charles-Edwards G, Glendenning J, Ganeshan B, Verma H, et al. Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy. Radiology. 2014 Jul;272(1):100–12.
118. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, et al. Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter Study [Internet]. Vol. 25, Clinical Cancer Research. 2019. p. 3538–47. Available from: http://dx.doi.org/10.1158/1078-0432.ccr-18-3190
119. Xiong Q, Zhou X, Liu Z, Lei C, Yang C, Yang M, et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy [Internet]. Clinical and Translational Oncology. 2019. Available from: http://dx.doi.org/10.1007/s12094-019-02109-8
120. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks [Internet]. Vol. 3, Journal of Medical Imaging. 2016. p. 034501. Available from: http://dx.doi.org/10.1117/1.jmi.3.3.034501
121. Li H, Giger ML, Huynh BQ, Antropova NO. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. J Med Imaging (Bellingham). 2017 Oct;4(4):041304.
122. Wong AJ, Kanwar A, Mohamed AS, Fuller CD. Radiomics in head and neck cancer: from exploration to application. Transl Cancer Res. 2016 Aug;5(4):371–82.

123. Toivonen J, Montoya Perez I, Movahedi P, Merisaari H, Pesola M, Taimen P, et al. Radiomics and machine learning of multisequence multiparametric prostate MRI: Towards improved non-invasive prostate cancer characterization. PLoS One. 2019 Jul 8;14(7):e0217702.

124. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O. Influence of segmentation margin on machine learning–based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas [Internet]. Vol. 29, European Radiology. 2019. p. 4765–75. Available from: http://dx.doi.org/10.1007/s00330-019-6003-8

125. Saini A, Breen I, Pershad Y, Naidu S, Knuttinen MG, Alzubaidi S, et al. Radiogenomics and Radiomics in Liver Cancers. Diagnostics (Basel) [Internet]. 2018 Dec 27;9(1). Available from: http://dx.doi.org/10.3390/diagnostics9010004

126. Badic B, Hatt M, Durand S, Jossic-Corcos CL, Simon B, Visvikis D, et al. Radiogenomics-based cancer prognosis in colorectal cancer. Sci Rep. 2019 Jul 5;9(1):9743.

127. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging [Internet]. 2019 Jun 18; Available from: http://dx.doi.org/10.1007/s00259-019-04372-x

128. Tian Y, Liu Z, Tang Z, Li M, Lou X, Dong E, et al. Radiomics Analysis of DTI Data to Assess Vision Outcome After Intravenous Methylprednisolone Therapy in Neuromyelitis Optic Neuritis. J Magn Reson Imaging. 2019 May;49(5):1365–73.

129. Bianchi J, Gonçalves JR, Ruellas AC de O, Vimort J-B, Yatabe M, Paniagua B, et al. Software comparison to analyze bone radiomics from high resolution CBCT scans of mandibular condyles. Dentomaxillofac Radiol. 2019 Sep;48(6):20190049.

130. Panth KM, Leijenaar RTH, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells [Internet]. Vol. 116, Radiotherapy and Oncology. 2015. p. 462–6. Available from: http://dx.doi.org/10.1016/j.radonc.2015.06.013

131. Leijenaar RT, Bogowicz M, Jochems A, Hoebers FJ, Wesseling FW, Huang SH, et al. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. Br J Radiol. 2018 Jun;91(1086):20170498.

132. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RTH, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. Oral Oncol. 2017 Aug;71:150–5.

133. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors [Internet]. Vol. 50, Journal of Nuclear Medicine. 2009. p. 122S – 150S. Available from: http://dx.doi.org/10.2967/jnumed.108.057307

134. Fukukita H, Senda M, Terauchi T, Suzuki K, Daisaki H, Matsumoto K, et al. Japanese guideline for the oncology FDG-PET/CT data acquisition protocol: synopsis of Version 1.0 [Internet]. Vol. 24, Annals of Nuclear Medicine. 2010. p. 325–34. Available from: http://dx.doi.org/10.1007/s12149-010-0377-7

135. Boellaard R, O'Doherty MJ, Weber WA, Mottaghy FM, Lonsdale MN, Stroobants SG, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0 [Internet]. Vol. 37, European Journal of Nuclear Medicine and Molecular Imaging. 2010. p. 181–200. Available from: http://dx.doi.org/10.1007/s00259-009-1297-4

136. Kinahan PE, Perlman ES, Sunderland JJ, Subramaniam R, Wollenweber SD, Turkington TG, et al. The QIBA Profile for FDG PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy. Radiology. 2020 Jan 7;191882.

137. Mankoff DA. Quantitative imaging as cancer biomarker [Internet]. Medical Imaging 2015: Physics of Medical Imaging. 2015. Available from: http://dx.doi.org/10.1117/12.2085907

138. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification [Internet]. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 2018. Available from: http://dx.doi.org/10.1109/isbi.2018.8363576

139. Mahmood F, Chen R, Durr NJ. Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training. IEEE Trans Med Imaging. 2018 Dec;37(12):2572–81.

140. Wang Q, Milletari F, Nguyen HV, Albarqouni S, Jorge Cardoso M, Rieke N, et al. Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings. Springer Nature; 2019. 254 p.

141. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. Int J Radiat Oncol Biol Phys. 2017 Oct 1;99(2):344–52.

142. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol. 2017 Jun;4:24–31.

143. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform. 2018 Apr;112:59–67.

144. Sheller MJ, Anthony Reina G, Edwards B, Martin J, Bakas S. Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation [Internet]. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. 2019. p. 92–104. Available from: http://dx.doi.org/10.1007/978-3-030-11723-8_9

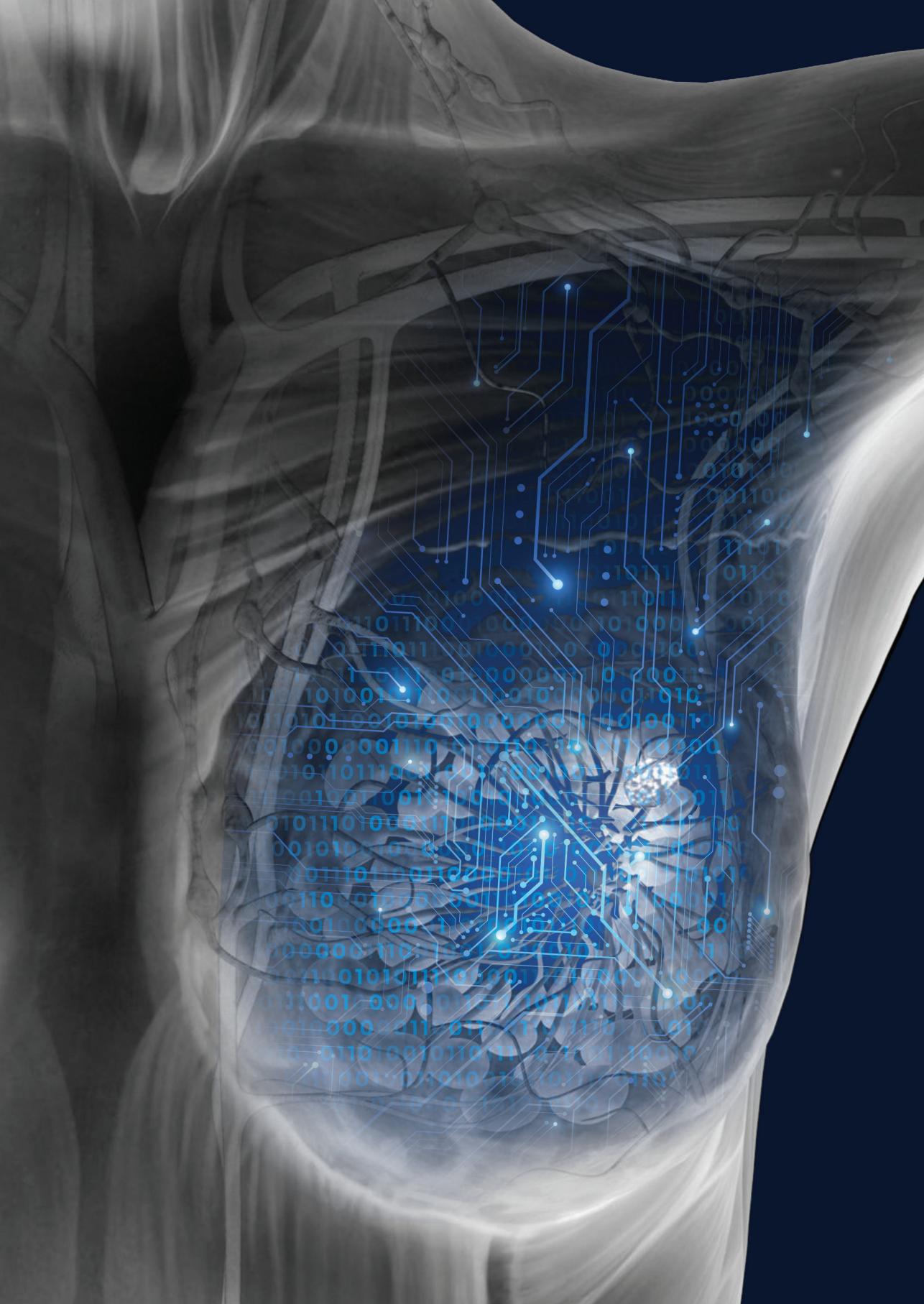
145. Lugan S, Desbordes P, Tormo LXR, Legay A, Macq B. Secure Architectures Implementing Trusted Coalitions for Blockchained Distributed Learning (TCLearn) [Internet]. 2019 [cited 2019 Oct 17]. Available from: http://arxiv.org/abs/1906.07690

146. Holzinger A. Explainable AI (ex-AI) [Internet]. Vol. 41, Informatik-Spektrum. 2018. p. 138–43. Available from: http://dx.doi.org/10.1007/s00287-018-1102-5

147. Khedkar S, Subramanian V, Shinde G, Gandhi P. Explainable AI in Healthcare [Internet]. SSRN Electronic Journal. Available from: http://dx.doi.org/10.2139/ssrn.3367686

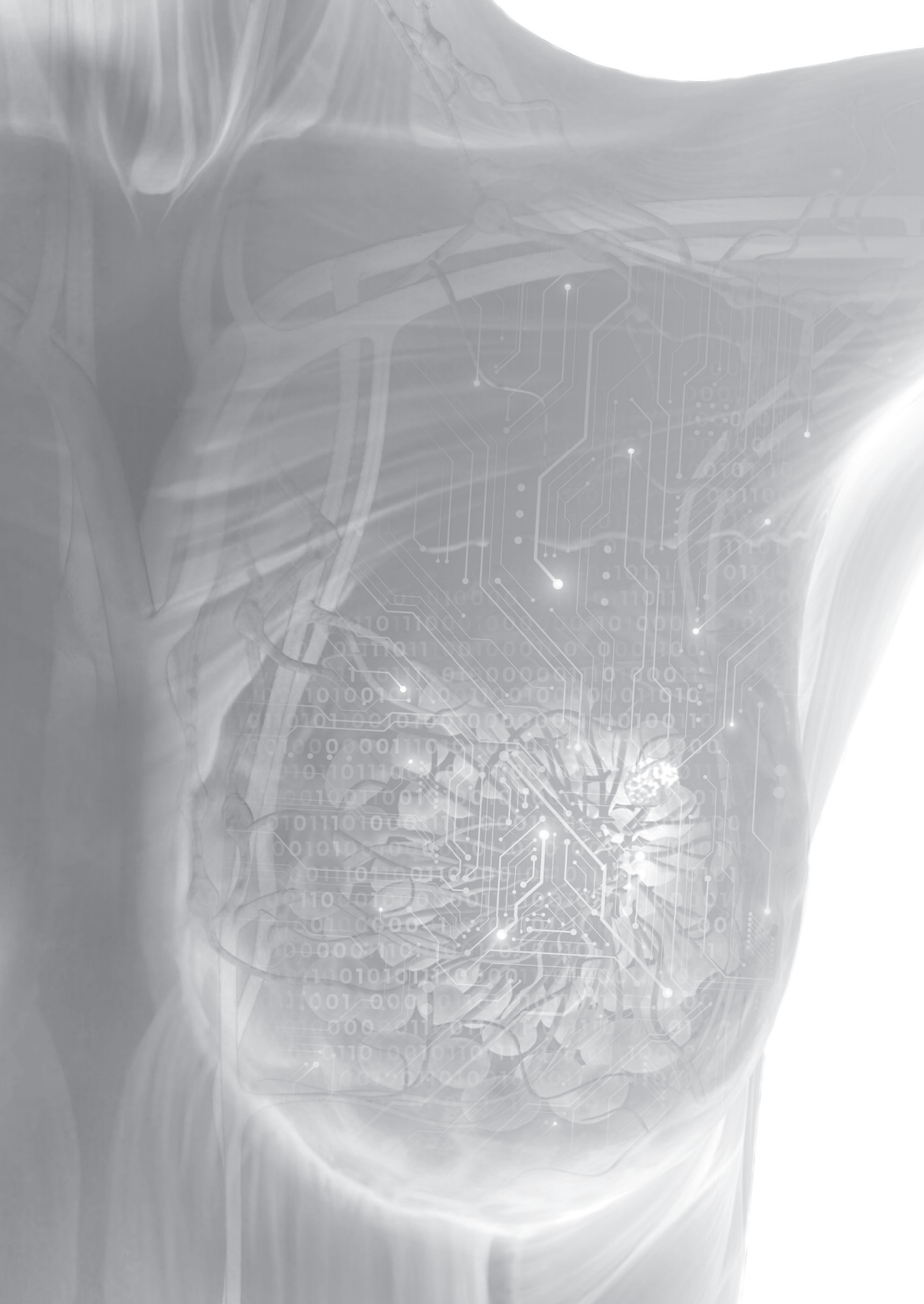
148. Joyner MJ, Paneth N. Promises, promises, and precision medicine [Internet]. Vol. 129, Journal of Clinical Investigation. 2019. p. 946–8. Available from: http://dx.doi.org/10.1172/jci126119

149. Saracci R. Epidemiology in wonderland: Big Data and precision medicine. Eur J Epidemiol. 2018 Mar;33(3):245–57.



PART I

MRI-based radiomics for prediction purposes in breast cancer patients



CHAPTER 3

Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: a systematic review

Renée W.Y. Granzier, Thiemo J.A. van Nijnatten, Henry. C. Woodruff, Marjolein L. Smidt, Marc B.I. Lobbes

Eur J Radiol. 2019 Dec;121:108736

Abstract

Background and purpose: MRI-based tumor response prediction to neoadjuvant systemic therapy (NST) in breast cancer patients is increasingly being studied using radiomics with outcomes that appear to be promising. The aim of this study is to systematically review the current literature and reflect on its quality.

Methods: A systematic literature search was performed using PubMed and EMBASE databases until May 8th, 2019. Abstracts were read and screened by two reviewers independently. The quality of the radiomics workflow of all eligible studies was assessed using the Radiomics Quality Score (RQS). An overview of the methodologies used in all steps of the radiomics workflow and current results are presented.

Results: Sixteen studies were selected for inclusion with cohort sizes ranging from 35 to 414 patients. The RQS scores varied from 0% to 41.2%. Methodologies used in the radiomics workflow varied greatly, especially for region of interest (ROI) segmentation, features selection, and model development with heterogeneous outcomes as a result. Seven studies applied univariate analysis and nine studies applied multivariate analysis. The majority of the studies performed their analysis on the pretreatment dynamic contrast-enhanced T1 weighted (DCE T1W) sequence. Entropy was the best performing individual feature with AUC values ranging from 0.83 to 0.85. The best performing multivariate prediction model, based on logistic regression analysis, scored an AUC of 0.94 in the validation cohort.

Conclusion: This systematic review revealed large methodological heterogeneity for each step of the MRI-based radiomics workflow, consequently, the (overall promising) results are difficult to compare. Consensus for standardization of MRI-based radiomics workflow for tumor response prediction to NST in breast cancer patients is needed to further improve research in this field.

Introduction

Neoadjuvant systemic therapy (NST) is increasingly used for breast cancer treatment [1-4]. Compared to adjuvant chemotherapy, NST bears the advantages of observing in vivo tumor response, decreasing tumor size (i.e., enabling breast-conserving therapy where initially mastectomy was indicated), and the possibility of achieving pathologic complete response (pCR) [5,6]. Tumor pCR is clinically relevant as it correlates with better prognosis expressed in improved disease-free survival and overall survival [7-9]. Unfortunately, not all patients respond well to NST, with treatment response ranging from pCR to non-response, and even to disease progression.

Different imaging modalities can be used to assess and predict tumor response to NST in breast cancer patients. Breast magnetic resonance imaging (MRI) is considered the most accurate imaging modality for both tumor assessment and response prediction [10,11]. However, its accuracy in assessing and predicting tumor response to NST remains insufficient to adapt treatment in clinical practice[12-14]. In addition, it is not possible to predict tumor response to NST based solely on the pretreatment MRI. Therefore, there is a continuous need to further improve breast MRI accuracy for this purpose.

Recent developments in tumor response prediction to NST show promising results in objectively interpreting MR images (usually from pre- and mid-treatment exams) using quantitative imaging analysis, such as radiomics. Radiomics is a high-throughput quantitative image analysis method in which standard-of-care medical images are converted into data that can be used to train machine learning models [15] (Figure 1). It has the advantages of detecting and quantifying the underlying structural heterogeneity of the entire breast tumor [16,17] at a level of detail far higher than can be achieved by visual assessment by radiologists.

The radiomics workflow consists of several consecutive steps. The methodology employed in all steps determines the quality of the final model. Different methodologies can lead to heterogeneous results which are difficult to compare. To the best of our knowledge, the variation in methodologies used in MRI-based radiomics for tumor response prediction to NST in breast cancer had not been evaluated before. This systematic review aims to assess the quality of the radiomics workflow using the Radiomics Quality Score (RQS) [18] and to report on the different methodologies used in all consecutive steps of the radiomics workflow. Furthermore, we summarize the current results reported on the topic of MRI-based radiomics for tumor response prediction to NST in breast cancer patients.

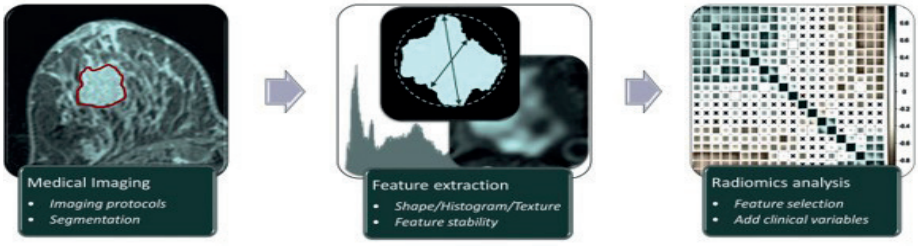


Figure 1. The radiomics workflow

Methods

Literature search

For this systematic review, a structured search using the PubMed and EMBASE (OVID) databases was conducted. The search was performed until May 8, 2019, and had a start date limit of January 1, 2010, approximately when the term 'radiomics' was introduced in medical publications [19]. Two researchers (R.G. and T.v.N.) independently performed the literature search to select potential studies.

Eligibility criteria & study selection

Studies were included if they met the following criteria: 1) The cohort consists of breast cancer patients who underwent at least a pre-treatment breast MRI; 2) Patients were treated with NST; 3) Breast MRI Radiomics analysis was used to predict tumor response to NST. 4) Studies are reported in English with institutional full-text availability. We excluded reviews, technical reports, letters to editors, comments to published studies, conference proceedings, as well as duplicate studies. When identical datasets were selected, the study reporting the most radiomics workflow details was chosen. The reviewers read all titles and abstracts independently and rejected studies that did not meet the aforementioned criteria. The full text of the remaining studies was determined for further eligibility by the same reviewers. In the case of discrepancies, a third reviewer (H.W.) was consulted to reach final consensus. For each included study, reference lists were searched manually for additional eligible studies.

Data extraction

The following data were extracted (if available): first author, year of publication, study design, analysis strategy, sample size, breast cancer subtype, NST regimen, MRI specifications (manufacturer, field strength, coil specifications), MR imaging parameters, MRI sequence, image pre-processing, feature extraction software, number of features, ROI segmentation methods (including profession and years of

those who performed segmentations), response definition, and results (including feature selection method, chosen classifier and diagnostic performance in terms of Area Under the receiver operating characteristic Curve (AUC)).

Data analysis

To quantify the radiomics workflow quality and its reporting, the included studies were evaluated in consensus by the two initial reviewers using the RQS. The RQS uses a checklist comprised of 16 components in the radiomics workflow, with 36 points (100%) representing a maximum score [18]. Descriptive analysis of the methodologies used in the included studies was performed according to the consecutive steps of the radiomics workflow: 1) image acquisition and preprocessing, 2) region of interest (ROI) segmentation, 3) feature extraction and selection 4) feature analysis and modeling, and 5) performance evaluation [15,20]. The response prediction results will be summarized in steps four and five of the radiomics workflow in the result section.

Results

Study selection

A total of 208 records were identified through the searches in PubMed (113 records) and EMBASE (95 records). After removing duplicates, the titles and abstracts of 155 records were screened, resulting in twenty-nine records eligible for inclusion. The full text of these studies was read and the selection criteria were applied, yielding a total of sixteen studies to be included in this systematic review. Figure 2 details the Preferred Reporting Item for Systematic reviews and Meta-Analyses (PRISMA) flow diagram, including the different screening phases.

General characteristics of included studies

The RQS heat map summarized the overall RQS scores which were considered as 'poor' with a mean score of 11.8% (range 0%- 41.2%), the most recently published studies scored the highest (Figure 3). A total of 1636 patients were retrospectively included in the sixteen studies presented in this review [21-36] (Table 1). Cohort sizes ranged from 35 to 414 patients, with a largest validation cohort of 137 patients. All breast cancer subtypes were included in all studies with the exception of Banerjee et al [25], which included triple-negative (TN) breast tumors only. To assess tumor response to NST, radiologic (RECIST) and pathologic methods were used. The studies that based their tumor response on pathology used varying definitions of response. While the majority of studies only used patients with no residual cancer burden as 'responders', four studies added patients with invasive residual cancer burden to that group [23,29,35,36].

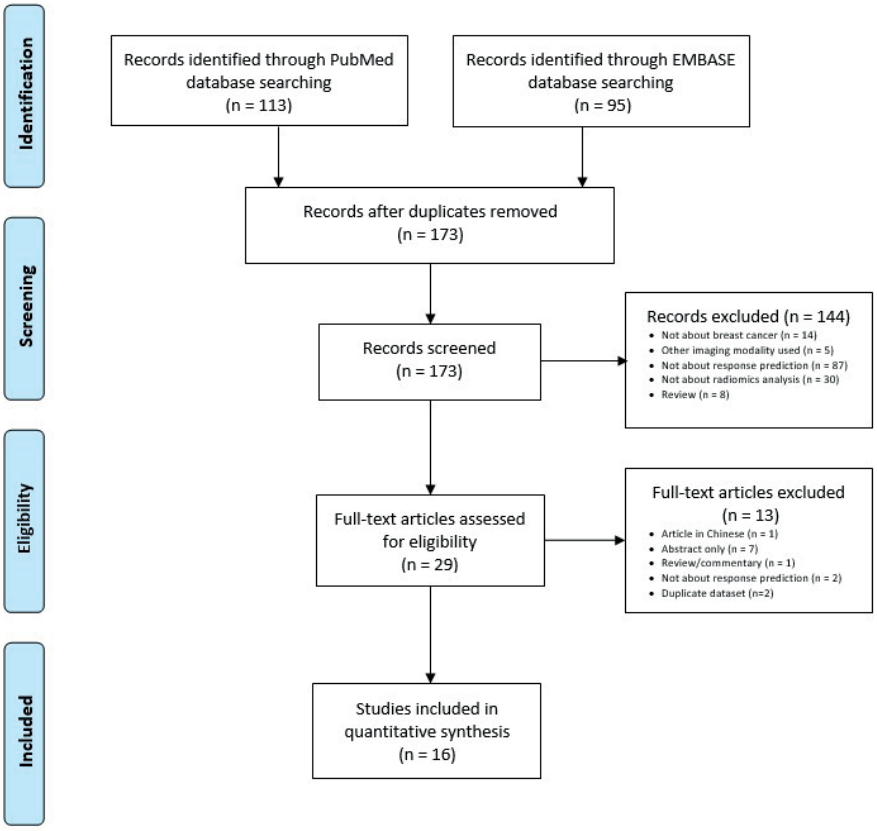


Figure 2. Flow diagram of preferred reporting items for systematic reviews and meta-analyses (PRISMA)

Radiomics workflow

Image acquisition and preprocessing

The authors of all studies, except Wu et al [32], performed their analysis on the pretreatment MRI. Two studies [28,34] analyzed the differences between MRI exams at two-time points. Both 1.5 and 3.0 Tesla MR scanners, with dedicated breast receiver coils, were used. The majority of the studies used the dynamic contrast-enhanced T1 weighted (DCE T1W) sequence [21,22,24,25,27,29-33,35,36], three studies T2W-sequences [26,28,34], and one study diffusion-weighted imaging sequence (DWI, using b-values of 0 and 1000 s/mm^2) [23]. Two studies applied multiparametric MRI for their analysis, consisting of DWI, DCE T1W, and T2W [21,22]. To exclude inter-scan differences one study used healthy contralateral breast parenchyma [28]. Wu et al [32] performed intratumor partitioning and Fan et al [29] added breast background parenchyma to their analysis region. Half of the studies reported on image pre-

processing performed by image discretization with histogram normalization [22,25,27,28,32,35,36]. Two studies performed only feature normalization[21,30].

RQS Component / Study	Image protocol quality (2)	Multiple segmentation (1)	phantom study (1)	Test-retest (imaging at multiple time points) (1)	Feature reduction or adjustment multiple testing (3)	Multivariate analysis with non radiomics features (1)	Biological correlates (1)	Cut-off analyses (1)	Discrimination statistics 92)	Calibration statistics (2)	Prospective study (7)	Validation (5)	Comparison to 'gold standard' (2)	Potential clinical utility (2)	Cost-effectiveness (1)	Open science and data (4)	Total score
Xiong (21)	1	0	0	0	3	1	0	1	2	1	0	2	2	0	0	0	**36.1%**
Liu (22)	1	1	0	0	3	1	0	0	2	0	0	5	2	0	0	0	**41.2%**
Yoon (23)	1	0	0	0	n/a	r/a	0	0	0	0	0	-5	2	0	0	0	**0.0%**
Hope Cain (24)	0	0	0	0	3	0	0	0	1	0	0	2	2	0	0	0	**19.4%**
Banerjee (25)	0	0	0	0	3	1	0	0	2	0	0	-5	2	0	0	0	**8.3%**
Chamming's (26)	1	0	0	0	n/a	r/a	0	0	1	0	0	-5	2	0	0	0	**0.0%**
Giannini (27)	1	0	0	0	3	0	0	0	2	0	0	-5	2	1	0	0	**11.1%**
Henderson (28)	1	0	0	0	n/a	r/a	0	0	1	0	0	-5	2	1	0	0	**0.0%**
Fan (29)	1	0	0	0	3	0	0	0	2	0	0	2	2	1	0	0	**30.6%**
Braman (30)	1	0	0	0	3	0	0	0	2	0	0	2	2	1	0	0	**30.6%**
Thibault (31)	1	0	0	1	n/a	n/a	0	0	1	[illegible]	0	-5	2	0	0	0	**0.0%**
Wu (32)	1	0	0	0	3	1	0	0	2	0	0	-5	2	0	0	0	**11.1%**
Michoux (33)	1	0	0	0	-3	1	0	0	2	0	0	-5	2	0	0	0	**0.0%**
Parikh (34)	1	0	0	0	n/a	n/a	0	0	1	0	0	-5	2	0	0	0	**0.0%**
Teruel (35)	1	0	0	0	n/a	n/a	0	0	1	0	0	-5	2	1	0	0	**0,0%**
Ahmed (36)	1	0	0	0	n/a	n/a	0	0	1	0	0	-5	2	0	0	0	**0.0%**

Figure 3. Heat map of radiomics quality score (RQS) per component of all included articles, n/a = not applicable

Region of interest segmentation

Six studies used 2D segmentations, five of which were performed manually [23,25,26,28,34], and one fully automatically [33]. Three studies used three adjacent representative slices with the maximum tumor diameter as their ROI, one performed manually [30], and two semi-automatically [35,36]. Seven studies performed 3D segmentation, three performed manually [21,22,31], three semi-automatically [24,29,32], and one automatically [27]. Different approaches were used for semi-automatic segmentation: 1) Point-and-click 3D segmentation, where segmentation started from a seed location determined by a mouse click. 2) Using a threshold to exclude necrotic, healthy, and/or fatty tissue followed by manual adjustments. 3) 3D box placing around the tumor after which an algorithm (fuzzy C-means) generated the tumor mask.

Segmentations were mostly performed by (breast) radiologists with 1 - 23 years of experience. Three studies did not report on this topic [23,28,36]. It remained unclear whether a radiologist performed the segmentations in the study of Wu et al [32] since they only reported on years of experience without naming a profession.

Feature extraction and selection

The number of features extracted ranged from 1 to 4650, mostly extracted with in-house built software. All studies, with the exception of four [26,28,32,34], extracted at least all Haralick textural features [37]. Feature selection only applied to the studies developing multivariate prediction models, where these studies all opted for a different approach, with regression methods chosen most often (Table 2). The study of Michoux et al [33] did not report on their feature selection method, and the study of Wu [32] did not perform feature selection at all. The number of features included in the models varied between two and twenty-four with a mean of six features.

Feature analysis and modeling

Over half of all studies presented individual features significant for tumor response prediction to NST (Table 2). Nine studies developed multivariate prediction, where most of these studies used a logistic regression model for the discriminant analysis [21,24,27,29,32,33]. Four studies used a support vector machine model (SVM). In addition to intratumoral feature extraction, the study of Braman et al [30] added a peritumoral region to the feature extraction region. The combination of peritumoral and intratumoral features yielded higher clustering accuracies for pCR prediction when compared to the feature groups individually (88% vs 71% vs 79%). Three studies [22,24,30] performed subgroup analyses for patients with different breast cancer subtypes, all showing better results in the models developed for the specific subtypes (Table 2). Henderson et al [28] showed excellent sensitivity and specificity for pCR prediction based on individual features within each subtype: ER+: 100%/ 100%; HER2+: 83.3%/95.7%, TN: 87.5%/80.0%. Furthermore, two studies explored individual features to distinguish TN breast cancer from all other subtypes [26,36]. Xiong et al [21] combined a clinical model based on HER2 and KI67 status, with a radiomics model. This combination showed improved results when compared to the radiomics model only (AUC of 0.94 vs 0.83). The best performing model developed by the study of Liu et al [22] also consisted of both clinical and radiomics features.

To validate the models, the majority of the studies performed leave-one-out cross-validation. Two studies divided their data into a training and a validation cohort [24,30], and two studies used an independent validation cohort from the same hospital [21,29]. Liu et al [22] was the only study to externally validate their model, using three external validation cohorts.

Performance evaluation

Entropy was the most common significant denominator with a highest AUC value of 0.85. In total, 26 different features showed significance. The AUC results of the multivariate analysis, for respective training and validation cohorts, ranged from 0.69 to 0.99 and 0.47 to 0.94. The multivariate model of the study of Xiong et al [21] showed overall the best performing classifier developed with an AUC value of 0.94 in the validation cohort. According to the published results of all radiomics models, no identical features could be identified (Table 2).

Discussion

This systematic review provides an overview of studies published on the topic of MRI-based radiomics for tumor response prediction to NST in breast cancer. Overall, the studies showed large methodological heterogeneity in each step performed along the radiomics workflow. Differences mostly arise due to the use of different ROI segmentation methods (2D vs. 3D), varying response definitions, and differences in modeling strategy. Despite the radiomics workflow heterogeneities, the results seem promising. Entropy was the most statistically significant individual feature reported. Validation results for articles reporting multivariate analysis ranged from AUC of 0.47 to 0.94. No identical features were identified in the multivariate analysis, where logistic regression was the most frequently chosen model.

In this review, we applied the RQS checklist to study the quality and reporting of the radiomics workflows. The studies employing multivariate analysis reached a mean score of 20.9% (0 – 41.2%). The studies using univariate analysis scored 0%, however, two items of the RQS (*i.e., feature reduction* and *multivariate analysis*) are not applicable to these studies, and hence no points were given. The RQS is also strongly influenced by the items *prospective research* and *validation cohort* since they account for respectively 7 and 5 points (*i.e.,* 33% of the maximum score). Further, the RQS score only identifies whether steps are being carried out and reported and not how they are performed. However, it is of utmost importance to know how the different methodologies used in each step of the radiomics workflow have been performed and applied, since differing methodologies and parameter settings can result in heterogeneous outcomes.

The first step in the radiomics workflow after image acquisition is ROI segmentation. This is an essential step since all subsequent steps rely on the (quality of the) segmentation. Segmentations are either performed manually (which is time-consuming, labor-intensive, and prone to inter-and intra-observer variation procedure), semi-automatically (which is prone to variation due to manual

contribution [38]) or automatically (which still bears the disadvantage of needing to be checked by experienced radiologists [39]). Recently, a study by Pandey et al [40] showed promising results with a fast and fully automatic MRI breast tumor segmentation software that reached accuracies above 95% comparable to manual segmentation. In this review, in ten studies segmentation was performed manually, five semi-automatically, and two automatically. In the publications using semi-automatic segmentation, four different methods were reported, which again indicates heterogeneity regarding tumor segmentation among the included studies.

Furthermore, the extent of the segmentation implementation is important for the results of the radiomics analysis. Previous research showed the advantages of 3D segmentation regarding measuring heterogeneity since feature extraction is performed on the entire tumor, one of radiomics most attractive features [41,42]. Of the studies included in this review, eleven used 3D segmentation, although four of these studies [30,35,36] only used a few consecutive slices, not fully exploiting this advantage of radiomics analysis. Standardization of a 3D segmentation method will enhance the applicability of future research to larger cohorts.

The next step in the workflow which also showed great variability was feature selection and model development. Feature selection was performed differently in all studies, with Michoux et al [33] not reporting on any feature selection details. Model overfitting seems to play a role in four studies [25,29,30,32] since the suggested need of at least ten patients for each feature in a model was exceeded [43]. It is important to exclude features that do not correlate with the outcome or highly correlate with other extracted features (e.g. volume) to reduce overfitting [20]**.** Model development was mostly performed by logistic regression. No identical features could be identified among the multivariate prediction models, to which the heterogeneity in previous steps probably contributed. A similar observation could be made for univariate feature analysis. The entropy feature demonstrated the most predictive ability for tumor response prediction to NST [23,27,28,33-35], but the lack of standardization of the preceding steps in the radiomics workflow did not allow for an overall conclusion.

The chosen outcome parameter for tumor response measurement varied among the studies. The majority of the studies used the commonly accepted gold standard: pathologic response assessment [44]. Only one study used the RECIST criteria based on imaging to quantify the response [29]. In the studies using pathologic assessment, response definitions varied, from no residual invasive cancer [22,24-28,30,31,34], to the consecutive addition of decrease in longest tumor diameter by >50% [36], of residual cancer burden below 1cm [35], or of any partial response [23]. The differences in outcome parameters made it impossible to compare results.

Furthermore, additional issues need to be addressed. First, most studies lumped all breast cancer subtypes into one category. Previous research has shown that tumor response prediction to NST varied among subtypes indicating the need for specific radiomics models, or at least the need to incorporate subtype into the analysis[8,45]. Radiomics models incorporating or differentiating for breast cancer subtypes will ensure more robust and comparable results. Second, the use of a multiparametric breast MRI approach adds additional information and yield more robust features [46]. Only the two most recently published studies [21,22] used this approach showing improved results. The majority of the studies opted for a single T1W contrast-enhanced imaging sequence [13]. However, a standard breast MRI protocol also consists of a T2W sequence and a DCE T1W sequence, sometimes combined with diffusion-weighted imaging and the derived apparent diffusion coefficient maps [47]. Future studies should further explore such multiparametric approaches, as they better reflect the clinical assessment performed by a radiologist. [48,49] Third, the lack of image pre-processing which is an important step in the radiomics workflow. Just over half of all studies applied and reported any method of image pre-processing. Image pre-processing ensures a more homogenous image quality (by interpolating all images to the same pixel size and slice spacing) [50,51]. Since feature values strongly depend on image quality it is of importance to perform this step, preferably in a standardized way. Fourth, publication bias, a more general limitation where studies with less favorable results may not be published while these may nevertheless contribute to the radiomics analysis. Lastly, since the field of radiomics is rapidly evolving, including the nomenclature, it might be possible that the search excluded eligible studies.

Twenty-six individual features showed significant results, with a mean AUC of 0.72 [range 0.65-0.85]. The mean AUC results for the multivariate models were 0.81 [range 0.70-0.99] and 0.72 [range 0.47-0.94] for the training and validation cohort, respectively. Despite encouraging results, similar to results from studies publishing on the same topic in different tumor sites 52,53, studies showed their concerns about clinical implementation. These concerns arose mostly from the methodological differences in the radiomics workflow, which also prevented us from performing a meta-analysis. Standardizing the methodologies used in the radiomics workflow will be the first step towards clinical implementation.

In this review, we aimed to perform a systematic overview of the methodologies used in the radiomics workflow and reported results in the field of MRI-based tumor response prediction to NST. Though limitations (including the heterogeneous radiomics workflow) currently plaguing this research field, results are still promising. Therefore we propose several recommendations which should be considered in designing future studies on radiomics research: (1) obliging image pre-processing performance, preferably in a standardized way; (2) consensus

should be reached for both ROI segmentation method and the (pathologic) response definition, or at least they should be described in detail; (3) automatic 3D segmentation of the tumor lesion to improve feature stability; (4) future models can be improved by incorporating breast cancer subtypes information and by using multiparametric MRI; (5) the application of external model validations to ensure that models are not simply reflecting localized spurious correlations between features and outcomes; (6) extensive reporting of each consecutive step in the radiomics analysis to increase transparency and reproducibility;

To conclude, studies focusing on MRI-based radiomics for tumor response prediction to NST in breast cancer patients showed promising results despite large methodological heterogeneity in each step of the radiomics workflow. This review demonstrates the requirement for more standardized methodology in the radiomics workflow since it is important to achieve robust and reproducible results in future research in order to translate the results to clinical applications.

Table 1. Overview of included studies in this review.

Author	Year	Study design	Sample Size	Analysis strategy	Software	MR Imaging parameters explained (yes/no)	MR Specification		
							Manufacturer	Field strength	Channels of breast Coil
Xiong [21]	2019	R	125	M	-	Yes	Philips, GE	1.5T, 3.0T	-
Liu [22]	2019	R	414	M	Built in-house Matlab toolbox	Yes	Philips	1.5T, 3.0T	-
Yoon [23]	2018	R	83	U	CGITA 1.3	Yes	Philips	3.0T	7
Hope Cain [24]	2018	R	288	M	-	No	Siemens, GE	1.5T, 3.0T	-
Banerjee [25]	2018	R	53	M	-	Yes	25 different sites	1.5T, 3.0T	-
Chamming's [26]	2018	R	85	U	TexRAD	Yes	GE	1.5T	8
Giannini [27]	2017	R	44	M	Built in-house	Yes	GE	1.5T	8
Henderson [28]	2017	R	88	U	MaZda 4.7	Yes	Siemens	3.0T	7
Fan [29]	2017	R	103	M	-	Yes	Siemens	1.5T, 3.0T	8
Braman [30]	2017	R	117	M	Built in-house	Yes	Philips*, Siemens*	1.5T, 3.0T	-
Thibault [31]	2017	R	38	U	-	Yes	Siemens	3.0T	4

Sequence	Phase DCE T1W used	Image pre-processing	Number of features extracted	Feature type (number)	Regimen neoadjuvant systemic therapy (n)	ROI segmentation method	Definition of responders
DCE T1W, T2WI, DWI	2	No*	647 (per sequence)	Geo(8), 1e(17), GLCM(22), GLRLM(14), NGTDM(5), GLSZM(13), wavelet	T (110) AC-T (15) *	3D, M	Non-responders[a]
DCE T1W, T2WI, DWI	2	N	4650 (per sequence)	Geo(8), 1e(17), T(99), wavelet(4535)	T (193) AC (63) T-AC (158)	3D, M	Complete response[b]
DWI	n/a	n/a n/a	46	IH(5), CM(6), VAM(11), NIDM(5), ISZM(11), TSM(2), NGLCM(6), NGLFM(5)	A-D (13) AC (60) H/Tr (10)	2D, M	Complete with partial response[d]
DCE T1W	1	No	529	Geo(15), T-E(30), FGT-E(82), T-E-T(135), FGT-E-T(135), T-E-V(35), FGT-E-V(34), SVTH(4), TFGT-E(18)	NS	3D, SA	Complete response[c]
DCE T1W, T2W	NS	N	538	GLCM(24), Riesz(72), Geo(20), IH(24), 1e(398)	G-C-I (53)	2D, M	Complete response[d]
T2W	1	No	6	IH(6)	AC-P (72) AC-D (4) EC-P (2) FEC-P (4) Ca-P (3)	2D, M	Complete response[b]
DCE T1W	1	N	27	GLCM(17), GLRLM(10)	A-P (28) A-P-Tr (14)	3D, A	Complete response[b]
T2W	n/a	N	1	E	FEC (18) FEC-D (41) FEC-D, Tr (26) D,P,TDM-1 (3)	2D, M	Complete response[e]
DCE T1W	Pre, 1, 2	No	158	IH(11), T(33), 1e(33), Dyn(84)	NS	3D, SA	Complete with partial response[f]
DCE T1W	1	No*	1980	T(99), PK (3), 1e(5)	A-T (89) D-Tr (28)	3D*, M	Complete response[g]
DCE T1W	2	No	1043	GLCM, GLRLM, GLSZM, LBP, PS, IH	AC-T (31) ISPY-2 (7)	3D, M	Complete response[e]

Author	Year	Study design	Sample Size	Analysis strategy	Software	MR Imaging parameters explained (yes/no)	MR Specification		
							Manufacturer	Field strength	Channels of breast Coil
Wu [32]	2016	R	35	M	-	Yes	Philips	3.0T	-
Michoux [33]	2015	R	69	M	Open-source Matlab codes	Yes	Philips	1.5T	-
Parikh [34]	2014	R	36	U	TexRAD	Yes	Siemens	1.5T	4 or 16
Teruel [35]	2014	R	58	U	In-house build	Yes	Siemens	3.0T	4
Ahmed [36]	2013	R	100	U	In-house build	Yes	GE	3.0T	8

Abbreviations: n/a = not applicable, - = unknown, NS = not specified
Study design: P = prospective, R = retrospective
Analysis strategy: U = Univariate feature analysis, M = multivariate prediction models
Sample size: *92 new patients compared to Braman[30]
MR Manufacturer: *three different Philips scanners, *three different Siemens scanners
Field strength: T = Tesla
Sequence: DCE-T1W = dynamic contrast-enhanced T1-weighted image, T2W = T2-weighted image, DWI = diffusion weighted image
Phase DCE T1W used: pre = pre-contrast, 1 = 1 minute post-contrast, 2 = 2 minutes post-contrast, 3 = 3 minutes post-contrast, 4 = 4 minutes post-contrast, 5 = 5 minutes post-contrast, *phases were used as separate models
Image pre-processing: N = normalization, No = not performed, * = features values were normalized.
Type of feature: E = Entropy, U = Uniformity, IH = intensity histogram, CM = co-occurrence matrix, VAM = voxel-alignment, GLCM = grey-level co-occurrence matrices, NIDM = neighborhood intensity difference matrix, ISZM = intensity size-zone matrix, NGLCM = normalized gray-level co-occurrence matrix, NGLDM = neighborhood gray-level dependence matrix, TSM = texture spectrum matrix, GLRLM = gray-level run length matrix, GLSZM = gray-level size zone matrix, LBP = local binary pattern, PS = pattern spectrum, Geo = geometric, T = texture, T-E-(T)= tumor enhancement (texture), FGT-E-(T) = fibroglandular tissue enhancement (texture), T-E-V = tumor enhancement variation, FGT-E-V = FGT-E-variation, SVTH = spatial variation of tumor heterogeneity, TFGT-E = combination of tumor and FGT enhancement, 1^e = first order statistics, Dyn = dynamic, K = kinetic, PK = pharmacokinetic

Sequence	Phase DCE T1W used	Image pre-processing	Number of features extracted	Feature type (number)	Regimen neoadjuvant systemic therapy (n)	ROI segmentation method	Definition of responders
DCE T1W	3	N	4	GLCM(4)	-	3D, SA	Complete response[d]
DCE T1W	2	No	25	K(3), Geo(2), GLCM(9), GLRLM(11)	A-T (43) A-T-Tr (26)	2D, M,A	Non-responders
T2W	n/a	No	2	E, U	FEC-D (7) EC-D (29) EC-D, Tr (6)	2D, M	Complete response[b]
DCE T1W	2	N	16	GLCM(16)	FEC (28) FEC-T (30)	3D*, SA	Complete with partial response[h]
DCE T1W	1,2,3,4 and 5*	N	16	GLCM(16)	EC-D (57) NS (38)	3D*, SA	Complete with partial response[k]

Regimen neoadjuvant systemic therapy: D = docetaxel, C = carboplatin, Tr = trastuzumab, PM = pertuzumab, A = doxorubicin, T = taxane, P = paclitaxel, AC = anthracycline, cyclophosphamide, H/Tr = hormone therapy or trastuzumab, G = gemcitabine, I = iniparib, EC = epirubicin- cyclophosphamide, FEC = fluorouracil, epirubicin- cyclophosphamide, Ca = carboplatin, TDM-1 = trastuzumab, emtansine, ISPY -2 = chemotherapy trial, *53 patients received trastuzumab

Delineation: M = manual, SA = semi-automatic, A = automatic, *three adjacent representative slices with largest tumor area

Definition of responders: [a]pCR defined as the absence of invasive cancer in the breast surgical specimen (ypT0/is), [b]Miller-Payne grade 1 and 2, [c]pCR defined as ypT0/isN0, [d]complete and partial response via Sataloff classification, [e]pCR defined as absence of invasive or in situ disease in the breast and/or lymph nodes (ypT0N0), [f]pCR defined via RCB scoring system; RCB = 0, [g]RECIST criteria, complete responder (CR) and partial responder (PR), [h] absence of any residual invasive cancer OR residual cancer burden below 1 cm, [k]absence of any residual invasive cancer OR a decrease in longest tumor diameter of greater than 50%

Table 2. Results of all included studies.

Author	Number of patients	Time point MRI exam	Analysis strategy	Univariate feature analysis				
				Significant features	p-value	AUC	Feature selection	Number of features
Xiong [21]	125	P	M	-	-	-	WRST, LASSO	4
Liu [22]	414	P	M	-	-	-	UFA, PCC, RFB	4
Yoon [23]	83	P	U	Entropy (histogram)	0.024	-	n/a	n/a
				ASM	0.033	-		
				Entropy	0.025	-		
Hope Cain [24]	288	P	M	-	-	-	MRB	2
							MRB	2
	151*	P	M	-	-	-	MRB	4
							MRB	
Banerjee [25]	53	P	M	-	-	-	-	*
								24
Chamming's [26]	85	P	U	Kurtosis (SFF = 2)	0.015	0.67	n/a	n/a
				Kurtosis (SFF = 4)	0.044			
				Kurtosis (SFF = 6)	0.019			

Multivariate feature analysis							
Model/ Classifier	**Model validation**	**AUC (training)**	**Accuracy training**	**Sens (%)**	**Spec (%)**	**AUC (Validation)**	**Accuracy validation**
MLR	TV	MRS: 0.92 (95% [0.84-1.00])	-	-	-	0.83 (95% [0.65-1.00])	-
		MCS: 0.99 (95% [0.96-1.00])	95.2% (93.4-97.1%)	-	-	0.94 (95% [0.85-1.00])	93.55% (91.44-95.69%)
SVM	EV*	MRS: 0.79 (95% [0.71-0.87])	-	-	-	0.79 (95% [0.65-0.93])	-
		MCS: 0.86 (95% [0.80-0.92])	-	-	-	0.80 (95% [0.67-0.91])	-
		HR+/HER2- 0.81 (95% [0.69-0.93])	-	-	-	0.87 (95% [0.66-1.00])	-
		HER2+ 0.70 (95% [0.56-0.85])	-	-	-	0.79 (95% [0.59-0.99])	-
		TN 0.96 (95% [0.89-1.00])	-	-	-	0.84 (95% [0.69-1.00])	-
n/a	n/a	-	-	-	-	n/a	n/a
LR	TV	-	-	-	-	0.66 (95% [0.56-0.76])	-
SVM	TV	-	-	-	-	0.59 (95% [0.48-0.70])	-
LR	TV	-	-	-	-	0.71 (95% [0.58-0.83])	-
SVM	TV	-	-	-	-	0.71 (95% [0.58-0.83])	-
LASSO	KFCV	0.69 ± 0.03	-	-	-	n/a	n/a
SVM	KFCV	0.74 ± 0.01	-	-	-	n/a	n/a
n/a	n/a	-	-	-	-	n/a	n/a

Author	Number of patients	Time point MRI exam	Analysis strategy	Univariate feature analysis				
				Significant features	p-value	AUC	Feature selection	Number of features
Giannini [27]	44	P	M	Contrast	<0.05	0.722	BRM	2
				Correlation	<0.05	0.715		
				Sum variance	<0.05	0.674		
				Difference variance	<0.05	0.699		
				Difference entropy	<0.05	0.719	FA	6
				LRE	<0.05	0.676		
				LRHGE	<0.05	0.708		
Henderson [28]	88	P, M	U	ΔEntropy fine	0.006	0.834	n/a	n/a
				ΔEntropy coarse	0.006	0.845		
Fan [29]	103	P	M	ΔBPE*	-	0.713	EAB	12
Braman [30]	117	P	M	-	-	-	mRMR	8
							mRMR	10
							mRMR	6
							mRMR	10
							mRMR	10
	70**	P	M	-	-	-	mRMR	10
	47***	P	<	-	-	-	mRMR	10
Thibault [31]	38	P, F	U	-	[c]	-	n/a	n/a
Wu [32]	35	F	M	Contrast	< 0.05	0.76	n/a	4
				Correlation	< 0.05	0.80		
				Energy	< 0.05	0.76		
				Homogeneity	< 0.05	0.76		

Multivariate feature analysis							
Model/ Classifier	**Model validation**	**AUC (training)**	**Accuracy training**	**Sens (%)**	**Spec (%)**	**AUC (Validation)**	**Accuracy validation**
LR	LOOCV	0.80 (95% [0.65-0.90])		80	69	n/a	n/a
Bayesian	LOOCV	-	0.70	67	72	n/a	n/a
n/a	n/a	-	-	-	-	n/a	n/a
LR	TV, LOOCV	0.91 (95% [0.80-1.00])	-	90	87.2	0.71 (95% [0.54-0.89])	-
LDA	TV, KFCV	0.75 ± 0.039	0.72	-	-	0.60	0.59
DLDA	TV, KFCV	0.78 ± 0.032	0.75	-	-	0.55	0.59
QDA	TV, KFCV	0.74 ± 0.037	0.76	-	-	0.64	0.64
Naïve Bayes	TV, KFCV	0.77 ± 0.021	0.78	-	-	0.69	0.64
SVM	TV, KFCV	0.71 ± 0.076	0.71	-	-	0.47	0.64
*DLDA	TV, KFCV	0.83 ± 0.025	0.79	-	-	-	-
Naïve Bayes	TV, KFCV	0.93 ± 0.018	0.84	-	-	-	-
n/a	n/a	-	-	-	-	n/a	n/a
LR	LOOCV	0.79 (95% [0.62-0.96])		75	78	n/a	-

Author	Number of patients	Time point MRI exam	Analysis strategy	Univariate feature analysis				
				Significant features	p-value	AUC	Feature selection	Number of features
Michoux [33]	69	P	M	Energy	0.002	0.702	-	2
				Entropy	0.003	0.696		
				Homogeneity	0.001	0.701		
				Inverse difference moment	0.001	0.711		
				Difference variance	0.023	0.649		4
				Run percentage	0.045	0.640		
				HGRE	0.038	0.644		
				Wash-in	0.008	0.685		
Parikh [34]	36	P, M	U	ΔEntropy	0.003	0.84	n/a	n/a
				ΔUniformity	0.004	0.84		
Teruel [35]	58	P	U	Sum variance	0.019	0.689	n/a	n/a
				Sum entropy	0.021	0.686		
				Entropy	0.040	-		
				Difference variance	0.040	-		
Ahmed [36]	100	P	U	Contrast	0.039[a]		n/a	n/a
				Difference variance	0.039[b]			

Abbreviations: n/a = not applicable, - = not performed
Number of patients: *triple negative/HER2+ subgroup analysis, **HR+/HER2- subgroup analysis, ***TN/HER2+ subgroup analysis.
Time point MRI exam: P = pretreatment, M = midtreatment, F = after first cycle of neoadjuvant systemic therapy
Significant features: ASM = angular second moment, SFF = spatial scaling factor, Δ between pretreatment and midtreatment, LRE = long run emphasis, LRHGE = long run high grey level emphasis, HGRE = high grey level run emphasis, *average difference of background parenchymal enhancement (BPE) between the breast with a breast tumor and the contralateral normal breast (best performing individual feature)
p-value: [a]1 min post-contrast, [b]2 min post-contrast, [c] Significant individual features could not be identified
Feature selection: WRST = Wilcoxon rank sum test, LASSO = least absolute shrinkage and selection operator, UFA = Univariate feature selection, PCC = Pearson correlation coefficient, RFB = Random forest Boruta, MRB = multilinear regression-based features selection, BRM = backward regression method, FA = filter approach, EAB = Evolutionary Algorithm-based feature selection, mRMR = minimum redundancy maximum relevance feature selection.

Multivariate feature analysis							
Model/ Classifier	**Model validation**	**AUC (training)**	**Accuracy training**	**Sens (%)**	**Spec (%)**	**AUC (Validation)**	**Accuracy validation**
LR	LOOCV	-	0.74	74	74	n/a	-
KMC	LOOCV	-	0.68	84	62	n/a	-
n/a	n/a	-	-	-	-	n/a	n/a
n/a	n/a	-	-	-	-	n/a	n/a
n/a	n/a	-	-	-	-	n/a	n/a

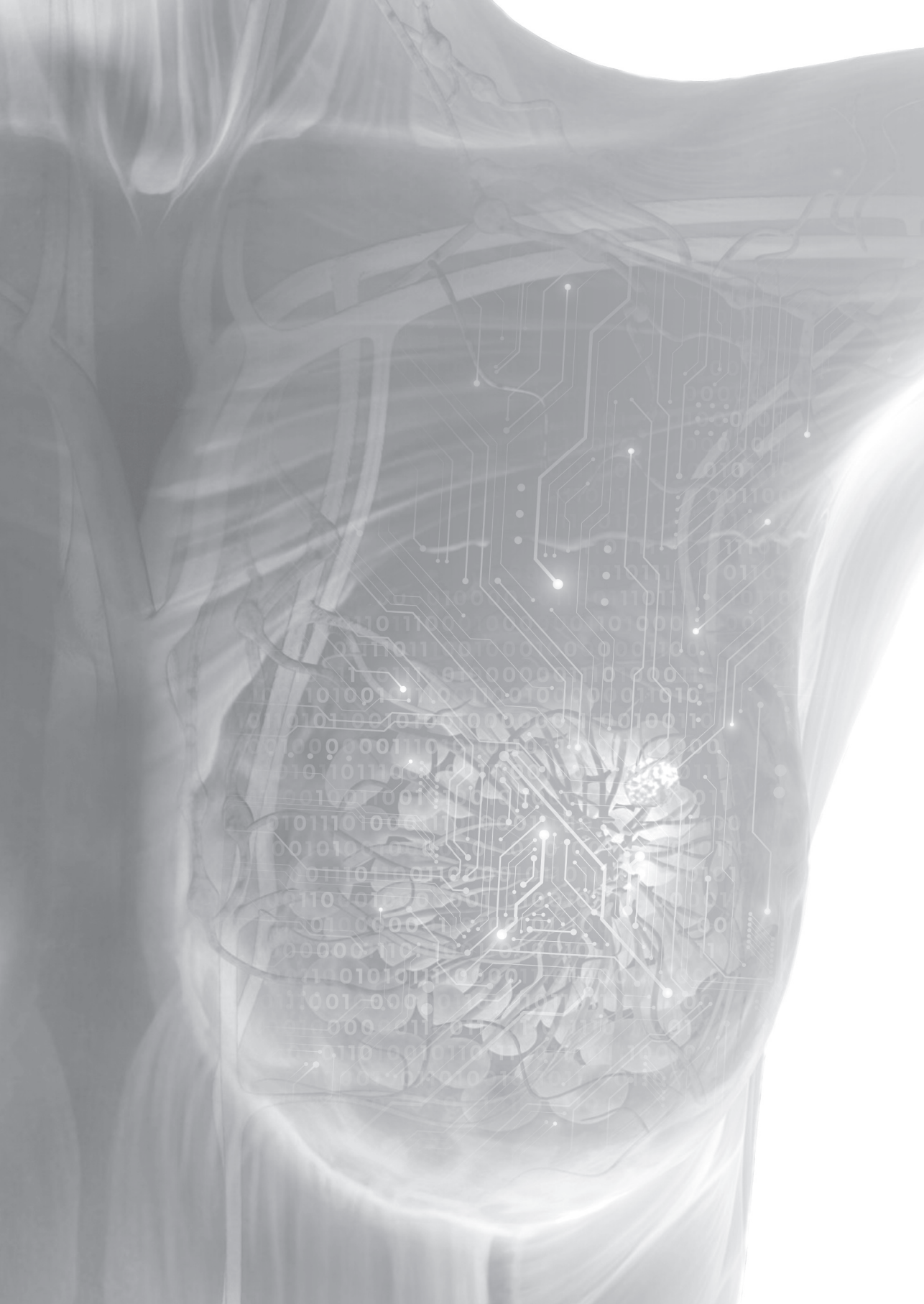
Number of features: ***unknown number of riesz and first-order features
Model/Classifier: MLR = multivariate logistic regression, SVM = support vector machine, LR = logistic regression, LDA = linear discriminant analysis, DLDA = diagonal linear discriminant analysis, QDA = quadratic discriminant analysis, KMC = k-means clustering, *best performing classifier.
Model validation: TV = cohort split in two; training and validation, EV = external validation, KFCV = k-fold cross validation, LOOCV = leave-one-out cross-validation.
AUC training: MRS = Multiparametric radiomics signature, MCS = Multiparametric combined signature (radiomics and clinical features), HR = hormone receptor, HER2 = Human epidermal growth factor receptor, TN = triple negative

References

1. Loibl S, Denkert C, von Minckwitz G. Neoadjuvant treatment of breast cancer--Clinical and research perspective. *Breast (Edinburgh, Scotland).* 2015;24 Suppl 2:S73-77.
2. Teshome M, Hunt KK. Neoadjuvant therapy in the treatment of breast cancer. *Surg Oncol Clin N Am.* 2014;23(3):505-523.
3. Spronk PER, de Ligt KM, van Bommel ACM, et al. Current decisions on neoadjuvant chemotherapy for early breast cancer: Experts' experiences in the Netherlands. *Patient Educ Couns.* 2018.
4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118.
5. Thompson AM, Moulder-Thompson SL. Neoadjuvant treatment of breast cancer. *Ann Oncol.* 2012;23 Suppl 10:x231-236.
6. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res.* 2015;21(2):249-257.
7. Bonnefoi H, Litiere S, Piccart M, et al. Pathological complete response after neoadjuvant chemotherapy is an independent predictive factor irrespective of simplified breast cancer intrinsic subtypes: a landmark and two-step approach analyses from the EORTC 10994/BIG 1-00 phase III trial. *Ann Oncol.* 2014;25(6):1128-1136.
8. Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet.* 2014;384(9938):164-172.
9. Prowell TM, Pazdur R. Pathological complete response and accelerated drug approval in early breast cancer. *The New England journal of medicine.* 2012;366(26):2438-2441.
10. Houssami N, Turner R, Morrow M. Preoperative magnetic resonance imaging in breast cancer: meta-analysis of surgical outcomes. *Annals of surgery.* 2013;257(2):249-255.
11. Hylton N. MR imaging for assessment of breast cancer response to neoadjuvant chemotherapy. *Magn Reson Imaging Clin N Am.* 2006;14(3):383-389, vii.
12. Lobbes MB, Prevos R, Smidt M, et al. The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: a systematic review. *Insights Imaging.* 2013;4(2):163-175.
13. Prevos R, Smidt ML, Tjan-Heijnen VC, et al. Pre-treatment differences and early response monitoring of neoadjuvant chemotherapy in breast cancer patients using magnetic resonance imaging: a systematic review. *Eur Radiol.* 2012;22(12):2607-2616.
14. Wasser K, Klein SK, Fink C, et al. Evaluation of neoadjuvant chemotherapeutic response of breast cancer using dynamic MRI with high temporal resolution. *Eur Radiol.* 2003;13(1):80-87.
15. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990).* 2012;48(4):441-446.
16. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006.
17. Davnall F, Yip CS, Ljungqvist G, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging.* 2012;3(6):573-589.
18. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749-762.
19. Gillies RJ, Anderson AR, Gatenby RA, Morse DL. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin Radiol.* 2010;65(7):517-521.

20. Ibrahim A, Vallières M, Woodruff H, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Seminars in Nuclear Medicine.* 2019.
21. Xiong Q, Zhou X, Liu Z, et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy. *Clin Transl Oncol.* 2019.
22. Liu Z, Li Z, Qu J, et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res.* 2019.
23. Yoon HJ, Kim Y, Chung J, Kim BSr. Predicting neo-adjuvant chemotherapy response and progression-free survival of locally advanced breast cancer using textural features of intratumoral heterogeneity on F-18 FDG PET/CT and diffusion-weighted MR imaging. *The breast journal.* 2018.
24. Cain EH, Saha A, Harowicz MR, Marks JR, Marcom PK, Mazurowski MA. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast cancer research and treatment.* 2018.
25. Banerjee I, Malladi S, Lee D, et al. Assessing treatment response in triple-negative breast cancer from quantitative image analysis in perfusion magnetic resonance imaging. *J Med Imaging (Bellingham).* 2018;5(1):011008.
26. Chamming's F, Ueno Y, Ferre R, et al. Features from Computerized Texture Analysis of Breast Cancers at Pretreatment MR Imaging Are Associated with Response to Neoadjuvant Chemotherapy. *Radiology.* 2018;286(2):412-420.
27. Giannini V, Mazzetti S, Marmo A, Montemurro F, Regge D, Martincich Lr. A computer-aided diagnosis (CAD) scheme for pretreatment prediction of pathological response to neoadjuvant therapy using dynamic contrast-enhanced MRI texture features. *Br J Radiol.* 2017;90(1077):20170269.
28. Henderson S, Purdie C, Michie C, et al. Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer. *Eur Radiol.* 2017;27(11):4602-4611.
29. Fan M, Wu G, Cheng H, Zhang J, Shao G, Li Lr. Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients. *European journal of radiology.* 2017;94:140-147.
30. Braman N, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res Journal Translated Name Breast Cancer Research.* 2017;19(1):no pagination.
31. Thibault G, Tudorica A, Afzal A, et al. DCE-MRI Texture Features for Early Prediction of Breast Cancer Therapy Response. *Tomography.* 2017;3(1):23-32.
32. Wu J, Gong GH, Cui Y, Li RJr. Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging.* 2016;44(5):1107-1115.
33. Michoux N, Van den Broeck S, Lacoste L, et al. Texture analysis on MR images helps predicting non-response to NAC in breast cancer. *BMC Cancer.* 2015;15:574.
34. Parikh J, Selmi M, Charles-Edwards G, et al. Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy. *Radiology.* 2014;272(1):100-112.
35. Teruel JR, Heldahl MG, Goa PE, et al. Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *NMR Biomed.* 2014;27(8):887-896.

36. Ahmed A, Gibbs P, Pickles M, Turnbull Lr. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *J Magn Reson Imaging.* 2013;38(1):89-101.
37. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics.* 1973;SMC-3(6):610-621.
38. Heye T, Merkle EM, Reiner CS, et al. Reproducibility of dynamic contrast-enhanced MR imaging. Part II. Comparison of intra- and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis. *Radiology.* 2013;266(3):812-821.
39. van Dam IE, van Sornsen de Koste JR, Hanna GG, Muirhead R, Slotman BJ, Senan S. Improving target delineation on 4-dimensional CT scans in stage I NSCLC using a deformable registration tool. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2010;96(1):67-72.
40. Pandey D, Yin X, Wang H, et al. Automatic and fast segmentation of breast region-of-interest (ROI) and density in MRIs. *Heliyon.* 2018;4(12):e01042.
41. Fave X, Cook M, Frederick A, et al. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Comput Med Imag Grap.* 2015;44:54-61.
42. Ng F, Kozarski R, Ganeshan B, Goh V. Assessment of tumor heterogeneity by CT texture analysis: Can the largest cross-sectional area be used as an alternative to whole tumor analysis? *European journal of radiology.* 2013;82(2):342-348.
43. Gillies R, Kinahan P, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology.* 2016;278(2):563-577.
44. Fowler AM, Mankoff DA, Joe BN. Imaging Neoadjuvant Therapy Response in Breast Cancer. *Radiology.* 2017;285(2):358-375.
45. von Minckwitz G, Untch M, Blohmer JU, et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *J Clin Oncol.* 2012;30(15):1796-1804.
46. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology.* 2018:181352.
47. Mann RM, Balleyguier C, Baltzer PA, et al. Breast MRI: EUSOBI recommendations for women's information. *Eur Radiol.* 2015;25(12):3669-3678.
48. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. *NPJ Breast Cancer.* 2017;3:43.
49. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep.* 2019;9(1):4800.
50. Lambin P. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol Journal Translated Name Nature Reviews Clinical Oncology.* 2017;14(12):749-762.
51. Jethanandani A, Lin TA, Volpe S, et al. Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review. *Front Oncol.* 2018;8:131.
52. Horvat N, Bates DDB, Petkovska I. Novel imaging techniques of rectal cancer: what do radiomics and radiogenomics have to offer? A literature review. *Abdom Radiol (NY).* 2019.
53. Chaddad A, Kucharczyk MJ, Daniel P, et al. Radiomics in Glioblastoma: Current Status and Challenges Facing Clinical Implementation. *Front Oncol.* 2019;9:374.



CHAPTER 4

MRI-based radiomics analysis for the pretreatment prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients: A multicenter study

Renée W. Y. Granzier, Abdalla Ibrahim[*], Sergey P. Primakov[*], Sanaz Samiei, Thiemo J. A. van Nijnatten, Maaike de Boer, Esther M. Heuts, Frans-Jan Hulsmans, Avishek Chatterjee, Philippe Lambin, Marc B. I. Lobbes, Henry C. Woodruff[†] and Marjolein L. Smidt[†]

[*]*Shared authorship*
[†]*Shared last authorship*

Cancers. 2021 May 18;13(10):2447

Abstract

This retrospective study investigated the value of pretreatment contrast-enhanced Magnetic Resonance Imaging (MRI)-based radiomics for the prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients. A total of 292 breast cancer patients, with 320 tumors, who were treated with neo-adjuvant systemic therapy and underwent a pretreatment MRI exam were enrolled. As the data were collected in two different hospitals with five different MRI scanners and varying acquisition protocols, three different strategies to split training and validation datasets were used. Radiomics, clinical, and combined models were developed using random forest classifiers in each strategy. The analysis of radiomics features had no added value in predicting pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients compared with the clinical models, nor did the combined models perform significantly better than the clinical models. Further, the radiomics features selected for the models and their performance differed with and within the different strategies. Due to previous and current work, we tentatively attribute the lack of improvement in clinical models following the addition of radiomics to the effects of variations in acquisition and reconstruction parameters. The lack of reproducibility data (i.e., test-retest or similar) meant that this effect could not be analyzed. These results indicate the need for reproducibility studies to preselect reproducible features in order to properly assess the potential of radiomics.

Introduction

Neoadjuvant systemic therapy (NST) is increasingly administered in the treatment of breast cancer. The number of breast cancer patients receiving NST varies between 17% and 70% and depends mainly on breast cancer subtype and tumor size[1,2]. NST allows monitoring of in vivo tumor response, potentially decreasing tumor size and thus enabling breast-conserving surgery[1,3,4]. Unfortunately, not all patients respond well to NST, with tumor response ranging from pathologic complete tumor response (pCR) to non-response and sometimes even progression of disease. Predicting which patients will respond well to NST and achieve tumor pCR could lead to modifications of treatment plans. In current clinical practice, magnetic resonance imaging (MRI) assessment combined with clinical (tumor) characteristics is used to determine tumor response to NST[5-7]. However, the diagnostic accuracy of the MRI with regard to tumor response evaluation is insufficiently accurate (76.1%) to adapt clinical treatment plans[8]. Furthermore, two studies investigated the use of ultrasound-guided biopsies to identify pCR after NST[9,10]. Unfortunately, the results showed that these biopsies are not accurate enough to identify pCR that surgery can be omitted[11].

Radiomics, a quantitative image analysis technique, could play a role predicting pCR from pretreatment dynamic contrast-enhanced (DCE)-MRI exams. Radiomics extracts large amounts of quantitative features from medical imaging, including MRI. These features capture information on the underlying heterogeneous structure of the region of interest (ROI), describing volume and shape, intensities and textures[12]. Radiomics' non-invasive ability to characterize the three-dimensional ROI, combined with the availability of ever-growing amounts of (longitudinal) imaging data and its cost-effectiveness, all contribute to the potential use of radiomics in personalized medicine[13-16]. The emergence of radiomics has so far mainly been applied in the field of clinical oncology and has also permeated breast cancer research.

Several MRI-based radiomics studies have reported promising results regarding the prediction of pCR to NST in breast cancer patients based on pretreatment scans[17-21]. However, the evidence from these studies is limited due to the relatively small sample sizes ranging from 29 to 100 patients and the lack of external validation datasets. Despite the promising potential of radiomics, several hurdles that impede the clinical implementation of radiomics models have been identified. One of these is the sensitivity of radiomics features to the variations in acquisition and reconstruction parameters across different imaging modalities[22-26], and some features were found not to be reproducible even in test-retest scenarios[27-29].

This study aimed to investigate the potential of pretreatment contrast-enhanced MRI-based radiomics for the prediction of pCR to NST in breast cancer patients.

We hypothesized that radiomics models trained and validated on data from two independent cohorts could add information to the prediction of tumor response to NST and that combined with clinical models can improve prediction accuracy. During our analysis, the sensitivity of radiomics features to the variations in acquisition and reconstruction parameters was established.

Materials and methods

Study population

In this multicenter study, imaging, and clinical data from consecutive women with histopathologically confirmed invasive breast cancer were retrospectively collected from two hospitals in the Netherlands (MUMC+—Maastricht University Medical Center and ZMC—Zuyderland Medical Center) between January 2011 and December 2018. The inclusion criteria were as follows: (i) treated with NST, (ii) have undergone pretreatment DCE-MRI in one of the two participating hospitals, and (iii) breast surgery after NST with histopathological outcome. Exclusion criteria were as follows: (i) histopathologically confirmed inflammatory breast cancer without the possibility of unequivocal tumor segmentation, (ii) MRI exam artefacts, if also rejected for visual assessment by the breast radiologist, (iii) non-standard chemotherapy regimens, deviating from the Dutch breast cancer guidelines, (iv) unfinished NST, and (v) no access to the patient's medical record. In the case of multifocal breast cancer, all histopathologically confirmed invasive tumors were included in the study. The institutional research board of both hospitals approved the study and waived the requirement for informed consent.

Study strategy

As different MRI scanners with varying acquisition and reconstruction parameters were used in the two hospitals, it was decided to develop separate prediction models (radiomics, clinical, and a combination of the two) for both cohorts and to validate them on each other (strategies 1 and 2). Therefore, all feature reduction, selection, and modeling procedures were performed on both data cohorts. A third modelling strategy was based on a mixture of both datasets divided into 70% training and 30% validation cohort. Feature selection and model building was performed on 70% of the training data and tested on the remaining 30% of the training data. The process of splitting the data into training and testing was iterated 100 times, maintaining class imbalance and ensuring that tumors from one patient were selected either in the training data or in the testing data. Figure 1A shows an overview of the selected data per strategy.

Clinical and pathological data

Clinical and pathological data were retrieved from patients' medical records and included age, clinical and pathological tumor, nodes, and metastases (TNM) stage, tumor grade, tumor histology, breast cancer subtype, and NST regimen. The majority of patients were treated with an anthracycline- and taxane-based NST regimen; the remaining received a taxane-based only NST regimen. Human epidermal growth factor receptor 2 (HER2) positive tumors received additional treatment with trastuzumab and/or pertuzumab. After completion of NST, all patients underwent breast surgery. The surgical specimens of all patients were evaluated via standard histopathological analysis by breast pathologists in the two participating hospitals. The breast tumor response was assessed by the Miller–Payne or Pinder grading systems[30,31]. In this study, tumors were defined as pCR when classified as grade 5 using the Miller–Payne classification or classified as 1i and 1ii using the Pinder classification (pCR; ductal carcinoma in situ may be present).

Imaging data

For all patients, pretreatment MRI exams were collected containing fat-suppressed 3D THRIVE DCE T1-weigthed (T1W), T2-weighted in the MUMC and fat-suppressed T2-weighted in the ZMC, and diffusion weighted imaging sequences. It was decided to only use the peak-enhanced phase of the DCE-T1W images for the radiomics analysis as tumors are best visible on this sequence[32,33]. The DCE-T1W images were obtained before and after intravenous injection of gadolinium-based contrast Gadobutrol (GadovistTM (EU)) with a volume of 15 mL and a flow rate of 2 mL/sec. A 105 s temporal resolution protocol was used in the MUMC+ and a 20 s temporal resolution protocol in the ZMC, resulting in five and nineteen post-contrast images for each patient in the MUMC+ and ZMC, respectively. Images were acquired using 1.5T (Ingenia, Intera, and Achieva by Philips Medical system and Avanto Fit by Siemens) and 3.0T (Skyra by Siemens) MRI scanners. All patients were scanned in prone-position using a dedicated breast-coil. DCE-T1W MRI acquisition protocols from both hospitals can be found in Table 1. Sequence parameters varied per MRI scanner and hospital, reflecting the heterogeneity in medical images used in daily clinical practice.

Tumor segmentation

The images acquired at tumor peak enhancement, at approximately two minutes' post-contrast administration, were used for the 3D ROI segmentation and further radiomics analysis, as tumors are best assessed on these images. All histologically confirmed invasive tumors were segmented manually using Mirada Medical's DBx 1.2.0.59 (64-bit, Oxford, UK) software by a medical researcher with three years of experience (RG), supervised by a dedicated breast radiologist with 14 years of experience (ML). During segmentation, the radiology reports were accessible,

and adjustment of image grayscale was allowed to optimize the visualization of the tumor. To gauge any bias introduced by inter-observer segmentation variability, 129 tumors from 102 patients acquired at MUMC+ were segmented by four observers independently with different degrees of experience in breast MR imaging (RG, ML, resident with three years of MRI experience (TvN), and a medical student with no experience (NV))[34].

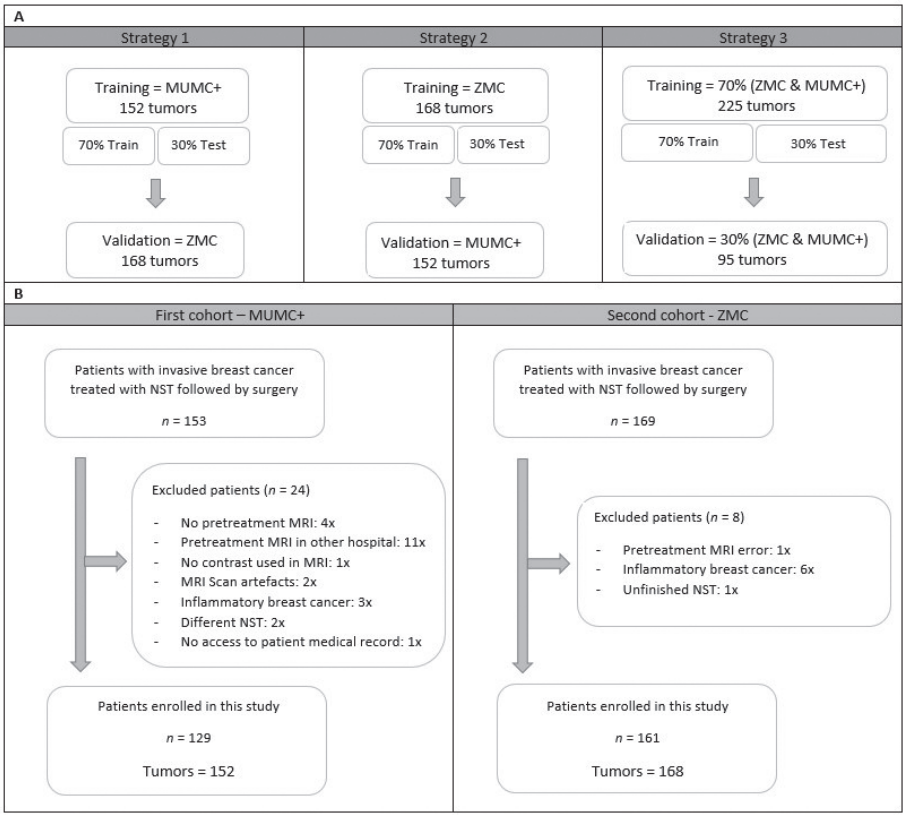


Figure 1. An overview of training, test, and validation data cohorts for the three strategies (A) and a flowchart from patient selection for the two different hospitals (B). Abbreviations, MUMC+ = Maastricht University Medical Center+, ZMC = Zuyderland Medical Center, NST = Neoadjuvant Systemic Therapy, MRI = Magnetic Resonance Imaging.

Table 1. Scanning Parameters.

Hospital	Scanner	Total MRI Exam No.	Group	No. of Tumors for Specific Scanning Parameters	Pixel Spacing	Acquisition Matrix (*n*)	Slice Thickness (mm)	TR/TE (ms) (*n*)	Spacing between Slices	Flip Angle
MUMC+	Philips 1.5T (Ingenia)	124	a	44	(0.97, 0.97)	340 × 340	1	3.4/7.5 3.5/7.6	1	10°
			b	66	(0.95, 0.95)	378 × 314 (28) 380 × 318 (23) 380 × 316 (18)	1	3.2/7.1 3.4/7.5 3.5/7.6	1	10°
			c	9	(0.80, 0.80)	344 × 344	1	3.4/7.5	1	10°
			d	3	(0.92, 0.92)	400 × 333 (2) 398 × 331 (1)	1	3.5/7.6 3.4/7.5	1	10°
			e	1	(0.88, 0.88)	384 × 368	1	3.4/7.5	1	10°
			f	1	(0.85, 0.85)	384 × 278	1	2.9/6.5	1	10°
	Philips 1.5T (Intera)	28	a	25	(0.97, 0.97)	340 × 337	1	3.4/7.4-7.6	1	10°
			b	1	(0.99, 0.99)	376 × 376	1	3.4/7.4	1	10°
			c	1	(0.95, 0.95)	364 × 364	1	3.4/7.5	1	10°
			d	1	(0.85, 0.85)	368 × 368	1	3.4/7.4	1	10°
ZMC	Philips 1.5T (Achieva)	123	a	94	(0.97, 0.97)	340 × 338	2	3.4/6.9-7.0	1	12°
			b	28	(0.96, 0.96)	372 × 368 (15) 372 × 370 (13)	2	3.4/6.9-7.0	1	12°
			c	1	(0.90, 0.90)	392 × 388	2	3.4/6.9	1	12°
	Siemens 3.0T (Skyra)	39	a	39	(0.69, 0.69)	288 × 288	2	1.2/4.0	unknown	10°
	Siemens 1.5T (Avanto_fit)	6	a	6	(0.89, 0.89)	224 × 202	2	2.4/6.1	unknown	10°

Abbreviations, MRI = Magnetic Resonance Imaging, TR = Repetition Time, TE = Echo Time, T = Tesla, MUMC+ = Maastricht University Medical Center+, ZMC = Zuyderland Medical Center.

Image pre-processing and feature selection

Image pre-processing of the two-minute postcontrast-T1W images was performed after tumor segmentation using an in-house developed pipeline and using a widely used proposed pre-processing method by Pyradiomics[35,36]. The in-house developed pipeline started first by applying bias field correction to every image using MIM software (version 6.9.4, Cleveland, Ohio, Unites States) to correct for nonuniform grayscale intensities in the MRI caused by field inhomogeneities. Second, in order to minimize acquisition-related radiomics variability, voxel dimensions were standardized across the cohorts to arrive at an isotropic voxel resolution of 1 mm3 by means of cubic interpolation[37]. Third, to homogenize arbitrary MRI units and clip image intensities to a certain range, a histogram matching technique was applied, adjusting the pixel values of the MR image such that its histogram matched that of the target MR image from the training data cohort[38-40]. Further gray value filtering was applied to generate MRIs with comparable gray value range and to enhance the contrast of the image using the following filtering parameters: window level (WL: 3050) and window width (WW: 2950). Filtering parameters were found when exploring the images after the histogram matching step. Fourth, to reduce high frequency noise and optimize handling of the image, grayscale values were resampled using a fixed bin width of 24, which reduced both image noise and computation times when extracting radiomics features from the ROI[41]. The pre-processing method proposed by Pyradiomics was applied after images' bias field correction and consisted of *z*-score normalization, resampling to isotropic voxel resolution of 1 mm3, and image discretization using a bin width of 100 to reach an ideal number of bins between 16 and 128[12].

For each ROI, 833 features were extracted using the Pyradiomics software (version 3.0). The extracted radiomics features included first-order statistics features (18), shape-based features (14), gray-level co-occurrence matrix features (GLCM) (22), gray-level run length matrix features (GLRLM) (16), gray-level size zone matrix features (GLSZM) (16), neighboring gray tone difference matrix features (NGTDM) (5), and gray-level dependence matrix features (GLDM) (14) from both unfiltered and filtered (eight wavelet decompositions) images.

Feature selection and radiomics model development

All feature selection steps followed by model development were performed on the 70% training data for each iteration. First, features sensitive to interobserver segmentation variabilities were removed using an intraclass correlation coefficient (ICC) cut-off value >0.75[29]. Consecutively, features with zero or small variance (with the frequency ratio between the most common value and the second most common value larger than 95/5) were removed. This was followed by the removal of highly correlated features using pairwise Spearman correlation (|r| >

0.90), where from any two highly correlated features, the feature with the highest mean correlation with the rest of the features was removed. Finally, the Boruta algorithm, a random forest feature selection method, was used to select important predictive features[42,43]. The Boruta algorithm duplicated all features and shuffled the values in the so-called shadow features. Random forest classifiers were trained on the real and shadow features, and the algorithm subsequently compared the importance score of each feature and selected only those features where the importance of the real feature was higher compared with the shadow's feature importance[44]. Random forest classification models were trained on the 70% of the training data and tested on the remaining 30% of the training data. The best performing radiomics models according to the summation of AUC and sensitivity value based on the test data in all strategies were selected and validated on the external validation data. All random forest parameters were set at default (Table S1) values. Figure 2 shows the radiomics workflow used in this study. Additionally, the range of the AUC values in the training data set is presented.

Clinical and combined model development

Clinical and combined (based on radiomics features and clinical variables) random forest models were trained, tested, and validated using the same strategy used to develop the radiomics models as described above. Clinical models were based on the available clinical characteristics, including age, clinical tumor stage (cT), clinical nodal stage (cN), clinical tumor grade, tumor histology, and breast cancer subtype. The best performing clinical and combined models according to the summation of AUC and sensitivity value based on the test data in all strategies were selected and validated on the external validation data. All random forest parameters were set as default. Additionally, the range of the AUC values in the training data set was presented.

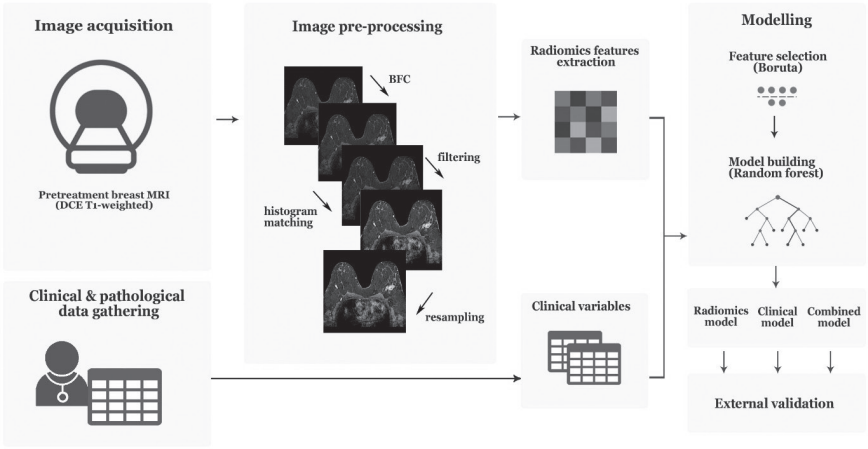


Figure 2. Radiomics workflow used in this study. Abbreviations, MRI = Magnetic Resonance Imaging, DCE = Dynamic Contrast-Enhanced, BFC = Bias Field Correction.

Statistical analysis

Image pre-processing steps were performed in Python (version 3.7) using an in-house developed pipeline based on the computer vision packages opencv (version 4.1.0), SimpleITK (version 1.2.0), and numpy (version 1.16.2) procedure. The remaining statistical analysis, feature selection, model development, and model evaluation were performed in R (version 3.6.3) using R studio (version 1.2.1335, Vienna, Austria)[45] and the R packages Boruta (version 7.0.0), Caret (version 6.0–85), Smotefamily (version 1.3.1), RandomForest (version 4.6–14), and pROC, (version 1.3.1)[46]. The difference between cohorts was assessed using independent samples t-test for continuous normally distributed variables, and Pearson chi-squared test for categorical variables. Statistical significance was based on p-values < 0.05 for both tests. The models developed were evaluated using the AUC and the 95% confidence interval (CI). DeLong's test was used to compare AUC values. In addition, the sensitivity and specificity and the negative predicted value (NPV) and positive predictive value (PPV) were derived from the confusion matrix. The radiomics quality score (RQS) was used to assess the radiomics workflow[14]. This study checked the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnoses (TRIPOD) guidelines[47,48].

Table 2. Clinical patient and tumor characteristics of patients in both complete data from the Maastricht University Medical Center+ (MUMC+) and Zuyderland Medical Center (ZMC) hospital.

Characteristics	MUMC+	ZMC	*p*-Value
Number of patients	129	161	-
Patient Age (years) (mean; range)	51 (28–73)	52 (28–79)	0.378
Number of tumors	152	168	-
Clinical tumor stage (%)			0.007
T1	29 (19.1)	16 (9.5)	
T2	99 (65.1)	103 (61.3)	
T3	20 (13.2)	37 (22.0)	
T4	4 (2.6)	12 (7.2)	
Clinical nodal stage (%)			<0.001
N0	88 (57.9)	59 (35.1)	
N1	44 (29.0)	87 (51.8)	
N2	9 (5.9)	12 (7.1)	
N3	11 (7.2)	7 (4.2)	
Unknown	0 (0.0)	3 (1.8)	
Clinical tumor grade (%)			0.003
1	8 (5.3)	22 (13.1)	
2	70 (46.1)	84 (50.0)	
3	68 (44.7)	62 (36.9)	
Unknown	6 (3.9)	0 (0.0)	
Tumor histology (%)			0.009
Invasive ductal carcinoma	136 (89.5)	134 (79.8)	
Invasive lobular carcinoma	10 (6.6)	14 (8.3)	
Invasive mixed ductal/lobular carcinoma	0 (0.0)	9 (5.4)	
Other invasive carcinoma	6 (3.9)	11 (6.5)	
Cancer Subtype (%)			0.921
HR+ and HER2−	80 (52.6)	82 (48.8)	
HR+ and HER2+	22 (14.5)	26 (15.5)	
HR− and HER2+	19 (12.5)	22 (13.1)	
Triple-negative	31 (20.4)	38 (22.6)	
Response to NAC (%)			0.331
pCR	53 (34.9)	49 (29.2)	
Non-pCR	99 (65.1)	119 (70.8)	

Abbreviations, HR = Hormone Receptor, HER2 = Human Epidermal growth factor Receptor 2.

Results

Patients demographics

A total of 322 women with invasive breast cancer and treated with NST were considered for inclusion, of whom 32 were excluded (Figure 1B). A total of 290 women with 320 breast tumors met the inclusion criteria, of whom 129 women with 152 breast tumors were collected at the MUMC+ and 161 women with 168 breast tumors at the ZMC. Table 2 summarizes the patient and tumor characteristics of both datasets. The pCR rate of the included tumors was 34.9% (53/152) and 29.2% (49/168) in the MUMC+ and ZMC cohorts, respectively, showing no significant difference. There were significant cohort differences in clinical tumor stage, clinical nodal stage, clinical tumor grade, and tumor histology. Clinical tumor stage, clinical tumor grade, and breast cancer subtype showed significant differences between pCR and non-pCR tumors within the individual cohorts (Table 3).

The results reported in the manuscript are based on the in-house developed image preprocessing pipeline, whereas the results based on the image pre-processing proposed by Pyradiomics are reported in the Supplementary Materials (Tables S2 and S3 and Figure S1). In both the radiomics and combined models, no significant differences were found (Table S4).

Radiomics models—feature selection and model performance

Of the 833 features extracted per ROI, 87 features were removed, as they were reported to be significantly affected by inter-observer segmentation variability (Table S5). In the best performing radiomics models in all strategies, one feature (*firstorder_maximum*) was removed, as it showed near zero variance. This was followed by the removal of: 574, 568, and 568 highly correlated features in strategy 1, 2, and 3, respectively, leaving 172, 178, and 178 features in the respective cohorts. The Boruta algorithm selected 5, 1, and 6 features in the best performing radiomics models for strategy 1, 2, and 3, respectively (Table 4A).

The results of the best performing radiomics models developed in the three strategies are shown in Table 5A. The AUC values in the validation cohorts were 0.55 (95% CI: 0.46–0.65), 0.52 (95%CI: 0.42–0.62), and 0.50 (95%CI: 0.37–0.64) for the respective strategies 1, 2, and 3. The sensitivity values ranged between 24% and 73% in the validation cohorts. The 100 radiomics models developed in the three strategies resulted in a range of AUC values in the training cohorts between 0.46 and 0.86 (Table S6).

Table 3. Clinical patient and tumor characteristics of patients in both complete data cohorts on pCR and non-pCR tumors from the Maastricht University Medical Center (MUMC+) and Zuyderland Medical Center (ZMC) hospitals.

Characteristics	MUMC+			ZMC		
	Non-pCR	pCR	*p*-Value	Non-pCR	pCR	*p*-Value
Number of tumors	99	53	-	119	49	-
Patient Age (years) (mean; range)	52 (32–72)	51 (28–73)	0.600	53 (31–79)	52 (28–73)	0.538
Clinical tumor stage (%)			0.019 *			0.023
T1	12 (12.1)	17 (32.1)		6 (5.0)	10 (20.4)	
T2	68 (68.7)	31 (58.5)		76 (63.9)	27 (55.1)	
T3	16 (16.2)	4 (7.5)		28 (23.5)	9 (18.4)	
T4	3 (3.0)	1 (1.9)		9 (7.6)	3 (6.1)	
Clinical nodal stage (%)			0.943			0.526
N0	56 (56.6)	32 (60.3)		39 (32.8)	20 (40.8)	
N1	29 (29.3)	15 (28.3)		62 (52.1)	25 (51.0)	
N2	6 (6.1)	3 (5.7)		11 (9.2)	1 (2.0)	
N3	8 (8.1)	3 (5.7)		5 (4.2)	2 (4.1)	
Unknown	0 (0.0)	0 (0.0)		2 (1.7)	1 (2.0)	
Clinical tumor grade (%)			<0.001 *			0.002
1	8 (8.1)	0 (0.0)		19 (15.9)	3 (6.1)	
2	58 (58.6)	12 (22.7)		66 (55.5)	18 (36.7)	
3	32 (32.3)	36 (67.9)		34 (28.6)	28 (57.2)	
Unknown	1 (1.0)	5 (9.4)		0 (0.0)	0 (0.0)	
Tumor histology (%)			0.913			0.030
Invasive ductal carcinoma	89 (89.9)	47 (88.7)		91 (76.5)	43 (87.8)	
Invasive lobular carcinoma	6 (6.1)	4 (7.5)		13 (10.9)	1 (2.0)	
Invasive mixed ductal/lobular carcinoma	0 (0.0)	0 (0.0)		9 (7.6)	0 (0.0)	
Other invasive carcinoma	4 (4.0)	2 (3.8)		6 (5.0)	5 (10.2)	
Cancer Subtype (%)			<0.001 *			<0.001
HR+ and HER2−	64 (64.6)	16 (30.2)		75 (63.0)	7 (14.3)	
HR+ and HER2+	15 (15.2)	7 (13.2)		14 (11.8)	12 (24.5)	
HR− and HER2+	6 (6.1)	13 (24.5)		5 (4.2)	17 (34.7)	
Triple-negative	14 (14.1)	17 (32.1)		25 (21.0)	13 (26.5)	

Abbreviations, pCR = pathologic Complete Response, HR = Hormone Receptor, HER2 = Human Epidermal growth factor Receptor 2.

Clinical models—feature selection and model performance

The clinical variables available were patient age, cT, cN, clinical tumor grade, tumor histology, and breast cancer subtype. None of the clinical variables were highly correlated. The Boruta algorithm selected four features in the best performing clinical models for all strategies (Table 4B). The results of the clinical models performed in the three settings are shown in Table 5B. The AUC values in the validation cohorts were 0.71 (95% CI: 0.62–0.79), 0.77 (95% CI: 0.70–0.85), and 0.72 (95% CI: 0.61–0.83) for strategy 1, 2, and 3, respectively. The clinical models performed significantly better compared with the radiomics models (Figure 3). The sensitivity values ranged between 41% and 47% in the validation cohorts. The 100 radiomics models developed in the three strategies resulted in a range of AUC values in the training cohorts between 0.68 and 0.88 (Table S6).

Combined models—feature selection and model performance

Of the 833 features extracted per ROI, 87 features were removed, as they were reported to be significantly affected by inter-observer segmentation variability. In the best performing combined models in all strategies, one feature (*firstorder_maximum*) was removed, as it showed near zero variance. This was followed by the removal of 580, 563, and 577 highly correlated features in strategy 1, 2 and 3, respectively, leaving 172, 189, and 175 features in the respective cohorts. The Boruta algorithm selected 7, 4, and 6 features in the best performing radiomics models for strategy 1, 2, and 3, respectively (Table 4C). The three models all contained the same clinical features, clinical tumor grade, and clinical breast cancer subtype. The results of the best performing combined models developed in the three strategies are shown in Table 5C. The AUC values in the validation cohorts were 0.73 (95% CI: 0.65–0.81), 0.69 (95%CI: 0.61–0.78), and 0.71 (95%CI: 0.60–0.81) for the respective strategies 1, 2, and 3. The sensitivity values ranged between 38% and 51% in the validation cohorts. The 100 radiomics models developed in the three strategies resulted in a range of AUC values in the training cohorts between 0.59 and 0.91 (Table S6).

RQS and TRIPOD results

This study scored a RQS score of 41.7% (15 out of 36 points) (Table S7). The score of the TRIPOD checklist was 73% (24 out of 33 applicable items).

Table 4. Selected features in best performing radiomics, clinical, and combined models for the three strategies.

	Strategy 1	Strategy 2	Strategy 3
A (Radiomics)	O_glszm_GrayLevelVariance	W.LHH_firstorder_Kurtosis	O_shape_Sphericity
	W.HLL_firstorder_Mean		W.LLH_glszm_GrayLevelNon-Uniformity
	W.HLL_glcm_Imc1		W.LLH_glszm_ZoneEntropy
	W.HLH_glcm_ InverseVariance		W.HHL_glcm_Imc1
	W.LLL_ngtdm_Complexity		W.HHH_glrlm_RunEntropy
			W.LLL_glcm_ DifferenceVariance
B (Clinical)	Age	cT	Age
	cT	cN	cT
	Tumor grade	Tumor grade	Tumor grade
	Breast cancer subtype	Breast cancer subtype	Breast cancer subtype
C (Combined)	Tumor grade	Tumor grade	cT
	Breast cancer subtype	Breast cancer subtype	Tumor grade
	O_shape_Sphericity	W.LHL_firstorder_kurtosis	Breast cancer subtype
	O_firstorder_Mean	W.HHL_gldm_ DependenceVariance	O_shape_Sphericity
	W.HLL_glcm_Imc2		W.LLH_glszm_SmallAreaL-owGrayLevelEmphasis
	W.HLL_glszm_ZoneEntropy		
	W.HLH_glcm_ InverseVariance		

Abbreviations: O = original, W = wavelet, cT = clinical tumor stage, and cN = clinical nodal stage.

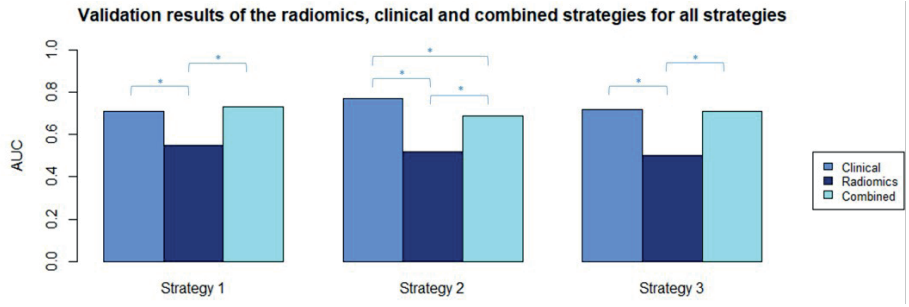


Figure 3. AUC values from the selected radiomics, clinical, and combined validation models in all strategies. * Significant difference between AUC values with p-value < 0.05 (p-values were calculated using the ROC test by Delong method).

Table 5. Performance of best performing random forest radiomics (5A), clinical (5B), and combined (5C) models for the three strategies.

A (Radiomics)	**Strategy 1**			**Strategy 2**			**Strategy 3**		
	Training MUMC+		**Validation ZMC**	**Training ZMC**		**Validation MUMC+**	**Training 70% Mixed**		**Validation 30% Mixed**
	Train	**Test**		**Train**	**Test**		**Train**	**Test**	
Area under the ROC	0.71	0.78	0.55	0.64	0.67	0.52	0.60	0.65	0.50
95% CI	0.59–0.82	0.63–0.92	0.46–0.65	0.54–0.75	0.49–0.84	0.42–0.62	0.49–0.71	0.51–0.80	0.37–0.64
Sensitivity (%)	53	59	73	44	60	28	38	48	24
Specificity (%)	89	79	36	75	72	62	92	77	88
PPV (%)	70	63	32	42	47	28	69	48	47
NPV (%)	79	76	77	77	81	62	75	77	72
B (Clinical)	**Strategy 1**			**Strategy 2**			**Strategy 3**		
	Training MUMC+		**Validation ZMC**	**Training ZMC**		**Validation MUMC+**	**Training 70% Mixed**		**Validation 30% Mixed**
	Train	**Test**		**Train**	**Test**		**Train**	**Test**	
Area under the ROC	0.79	0.81	0.71	0.81	0.84	0.77	0.75	0.86	0.72
95% CI	0.71–0.87	0.68–0.95	0.62–0.79	0.73–0.89	0.72–0.96	0.70–0.85	0.68–0.83	0.77–0.95	0.61–0.83
Sensitivity (%)	54	86	45	54	71	47	52	71	41
Specificity (%)	87	64	74	85	86	85	77	84	78
PPV (%)	69	57	42	59	67	63	52	68	46
NPV (%)	78	89	77	82	88	75	77	86	75

C (Combined)	Strategy 1			Strategy 2			Strategy 3		
	Training MUMC+		Validation ZMC	Training ZMC		Validation MUMC+	Training 70% Mixed		Validation 30% Mixed
	Train	Test		Train	Test		Train	Test	
Area under the ROC	0.82	0.83	0.73	0.79	0.86	0.69	0.79	0.86	0.71
95% CI	0.74–0.90	0.70–0.97	0.65–0.81	0.71–0.88	0.74–0.98	0.61–0.78	0.73–0.86	0.76–0.96	0.60–0.81
Sensitivity (%)	53	67	51	51	71	51	52	71	38
Specificity (%)	88	88	82	87	82	67	85	89	83
PPV (%)	69	77	53	62	63	45	61	75	50
NPV (%)	78	82	80	81	88	72	79	87	75

Abbreviations, MUMC+ = Maastricht University Medical Center+, ZMC = Zuyderland Medical Center, CI = confidence interval, PPV = positive predicted value, NPV = negative predicted value.

Discussion

In this multicenter study, we investigated the value of pretreatment contrast-enhanced MRI-based radiomics for the prediction of pCR to NST in breast cancer patients using radiomics, clinical, and combined models in three different data-mixing strategies. The AUC values of the radiomics, clinical, and combined models in the validation datasets of the three strategies had ranges of 0.50–0.55, 0.71–0.77, and 0.69–0.73, respectively. Different radiomics features were selected for the radiomics and combined models in the three strategies, while the selected clinical features were mostly the same in all scenarios, with comparable performances. These results indicate poor performance of the radiomics features and that the radiomic features had no added value to the clinical models developed for the prediction of pCR to NST in breast cancer patients.

The clinical models significantly outperformed the radiomics models for the prediction of pCR to NST in all strategies. This indicates that radiomics features in these scenarios did not have an added value to the clinical model we developed. Furthermore, the variation in the features selected and model performance was greater in the radiomics models compared with the clinical models. However, based on current knowledge in the radiomics field, we cannot say that radiomics features do not have an added value unless the variations in acquisition and reconstruction parameters are properly addressed. Due to the lack of reproducibility data, this study could not analyze the effects of different acquisition and reconstruction parameters on radiomics feature values. Furthermore, the significant differences in population characteristics between the two cohorts could have led to the low performance of the radiomics models. While there was overlap in breast cancer phenotypes, the proportions at which these phenotypes occur may have differed so that the differences in prevalence resulted in differences in overall classification performances.

The results of this study indicate that even extensive MRI pre-processing and homogenization of the MR images do not sufficiently address the variations in acquisition and reconstruction parameters. This is in line with studies published in recent years that investigated the reproducibility of MRI radiomics features in test-retest phantom data as well as in patient data of varying disease sites, and showed that, among others, the variations in acquisition and reconstruction parameters strongly influence the values (concordance) of radiomics features[24,27-29,49-52]. Shur et al.[29] performed a test-retest 1.5T MRI phantom study using the same imaging protocol and showed that 20% of the examined features were not repeatable. A study on repeatability and reproducibility using a T2W pelvic phantom showed that radiomics features values are not only affected by varying acquisition parameters but also by the use of different MRI vendors and magnetic field strengths, wherein

the reproducibility of the radiomic features is more affected by difference in MRI vendor than by difference in magnetic field strength[49]. Overall, they reported that only 3.3% (31/944) of the examined features showed excellent robustness (ICC and CCC > 0.9). The radiomics community is currently trying to address these major hurdles.

Investigating comparable published work, we found a number of studies using only univariate predictive features without an external validation data cohort[18-21,53,54] and more recent published papers that were focusing on multivariate prediction models[32,33,55,56]. Hope Cain et al.[55] achieved an AUC value of 0.71 (95% CI: 0.58–0.83) for predicting pCR to NST in TN/HER2+ breast cancer patients; however, the model was not externally validated. Therefore, we anticipate that the results could not be generalized to scans acquired with different vendors/parameters than those used in the study. The study by Liu et al.[57] was the only study performing external radiomics model validation for the prediction of pCR to NST in breast cancer patients. The study differed from our research by the use of multiparametric (T2-weighted, diffusion-weighted images, and contrast-enhanced T1-weighted) MRI. However, the use of multiple MRI sequences for pCR prediction achieved better outcome with validation AUC values between 0.71 and 0.80. However, it is remarkable that their external validation results were obtained with MRI images that were much less extensively pre-processed compared to our images.

Our study also has its limitations. First, selection bias in retrospective studies is inevitable and so are the biases introduced by clinical protocols, such as HER2+ tumors receiving additional treatment compared to other tumors. Second, since the effect of different MRI scanners and acquisition and reconstruction parameters on radiomics features in breast imaging is not determined, we could not adjust our model for the potential variance induced by these factors in the radiomics feature values. Therefore, since data were collected from two hospitals using five MRI scanners with different acquisition and reconstruction parameters, noise may have been introduced into the models by incorporating radiomics features not robust to these variations. Third, while we believe that MRI preprocessing is a necessary step toward comparable images with intensity values having similar tissue meaning, it is possible that with our choice of preprocessing steps, consistent with current literature, we may have inadvertently removed quantitative information. However, the results obtained with the widely used pre-processing method proposed by Pyradiomics showed no significant differences from the result reported here. Fourth, the number of patients included in this study did not allow us to perform a subanalysis for the different breast cancer subtypes. Fifth, the data were collected over a relatively long period of time during which optimization of MRI acquisitions protocols occurred, which may have introduced variations as well. Last, for these analyses it was specifically chosen

to use the peak-enhanced (2 min) post-contrast T1W images, as breast tumors are most visible on them and because some of the tumors included cannot be seen on other sequences; for example, mucinous tumors and some of the invasive lobular tumors are not or only weakly visible on the subtraction images. In our opinion, performing the analysis using the subtraction images instead of the peak-enhanced images would have resulted in a significant decrement in the number of patients that could be analyzed. Furthermore, as the effects of the different breast MRI sequences on the radiomics features is not yet understood, future radiomics research in the field of breast cancer could focus on the use of the different MRI sequences, as well as on multiparametric and delta radiomics approaches.

Conclusions

In conclusion, this study showed no contribution of pretreatment contrast-enhanced MRI-based radiomics for the prediction of tumor pCR on NST in breast cancer patients, as neither the radiomics nor the combined models performed significantly better than the clinical models. However, without analysis of the effects of variations in acquisition and reconstruction parameters, it is currently not possible to conclude that pretreatment contrast-enhanced MRI-based radiomic features have no value in the prediction of pCR to NST. The effects of different acquisition and reconstruction parameters on radiomics feature values in breast imaging should be explored in future MRI-breast reproducibility studies to investigate whether further research into pretreatment MRI-based radiomics for the prediction of pCR to NST in breast cancer patients is useful.

Supplementary materials

Table S1. Default random forest parameters.

Random Forest Parameters	Default Value
Ntree	500
Mtry	√(ncol(x))
Nodesizes	1
Maxnodes	NULL

The parameter *mtry* is the square root of the number of features to be included in the random forest model, which is equal to the square root of the number of columns (=ncol). Since the amount of included features differed per iteration we could not give a single number and therefor choose to give the equation. The parameter *maxnodes* has NULL as default value. This means that the maximum number of terminal nodes trees can grow to the maximum possible.

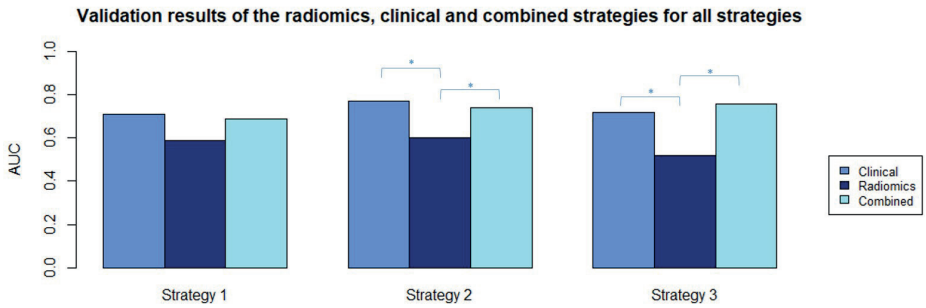


Figure S1. AUC values from the selected radiomics, clinical and combined validation models in all strategies. (P values were calculated using the roc test by Delong method).

Table S2. Selected features in best performing radiomics, clinical and combined models for the three strategies using the proposed imaging pre-processing method by Pyradiomics.

	Strategy 1	Strategy 2	Strategy 3
A (Radiomics)	O_shape_Sphericity	W.LHL_glcm_Correlation	O_shape_MajorAxisLength
	O_glszm_HighGrayLevel-ZoneEmphasis	W.HHH_glszm_ GrayLevelVariance	O_shape_Sphericity
	O_glszm_ZoneEntropy	W.LLL_firstorder_ Skewness	W.LLH_glrlm_RunEntropy
	W.LHL_glcm_Correlation		W.HLL_glszm_ZoneEntropy
			W.HLL_gldm_ DependenceEntropy
			W.HHH_glcm_Imc1
			W.HHH_glcm_ MaximumProbability
			W.LLL_firstorder_Skewness
			W.LLL_gldm_ LowGrayLevelEmphasis
B (Clinical)	Age	cT	Age
	cT	cN	cT
	Tumor grade	Tumor grade	Tumor grade
	Breast cancer subtype	Breast cancer subtype	Breast cancer subtype
C (Combined)	Tumor grade	Tumor grade	Age
	Breast cancer subtype	Breast cancer subtype	cT
	O_Shape_Sphericity	W.LHL_glcm_Idmn	Tumor Grade
	W.HLL_gldm_ DependenceEntropy	W.LLL_firstorder_ Skewness	Breast cancer subtype
	W.HLH_glcm_Imc2		W.LLH_glrlm_RunEntropy
	W.HHH_gldm_ DependencVariance		W.LHL_glcm_Imc1
			W.HLL_glszm_ZoneEntropy

Abbreviations, O = original, W = wavelet, cT = clinical tumor stage, and cN = clinical nodal stage

Table S3. Performance of best performing random forest radiomics (5A), clinical (5B) and combined (5C) models for the three strategies using the proposed imaging pre-processing method by Pyradiomics.

A (Radiomics)	Strategy 1			Strategy 2			Strategy 3		
	Training MUMC+		Validation ZMC	Training ZMC		Validation MUMC+	Training 70% Mixed		Validation 30% Mixed
	Train	Test		Train	Test		Train	Test	
Area under the ROC	0.69	0.69	0.59	0.65	0.70	0.60	0.70	0.61	0.52
95% CI	0.58–0.80	0.48–0.90	0.49–0.69	0.53–0.77	0.54–0.85	0.50–0.69	0.61–0.78	0.45–0.77	0.39–0.65
Sensitivity (%)	43	54	33	35	40	28	50	32	24
Specificity (%)	80	83	76	93	75	85	89	74	84
PPV (%)	55	58	36	67	40	50	69	33	41
NPV (%)	71	81	73	78	75	69	78	72	71

B (Clinical)	Strategy 1			Strategy 2			Strategy 3		
	Training MUMC+		Validation ZMC	Training ZMC		Validation MUMC+	Training 70% Mixed		Validation 30% Mixed
	Train	Test		Train	Test		Train	Test	
Area under the ROC	0.79	0.81	0.71	0.81	0.84	0.77	0.75	0.86	0.72
95% CI	0.71–0.87	0.68–0.95	0.62–0.79	0.73–0.89	0.72–0.96	0.70–0.85	0.68–0.83	0.77–0.95	0.61–0.83
Sensitivity (%)	54	86	45	54	71	47	52	71	41
Specificity (%)	87	64	74	85	86	85	77	84	78
PPV (%)	69	57	42	59	67	63	52	68	46
NPV (%)	78	89	77	82	88	75	77	86	75

C (Combined)	Strategy 1			Strategy 2			Strategy 3		
	Training MUMC+		Validation ZMC	Training ZMC		Validation MUMC+	Training 70% Mixed		Validation 30% Mixed
	Train	Test		Train	Test		Train	Test	
Area under the ROC	0.73	0.79	0.69	0.76	0.84	0.74	0.85	0.83	0.76
95% CI	0.64–0.83	0.65–0.94	0.60–0.77	0.66–0.86	0.71–0.96	0.66–0.82	0.79–0.91	0.73–0.93	0.65–0.86
Sensitivity (%)	49	69	24	49	64	58	60	62	45
Specificity (%)	76	70	85	88	88	76	86	84	86
PPV (%)	51	58	40	63	69	56	67	65	59
NPV (%)	74	79	73	81	86	77	82	83	77

Table S4. p-values for the comparison of the AUC values of the radiomics and combined models developed using the in-house developed image preprocessing (in-house) and the proposed image pre-processing by Pyradiomics (Pyradiomics).

In-house vs. Pyradiomics	Strategy 1	Strategy 2	Strategy 3
Radiomics	0.507	0.279	0.884
Combined	0.283	0.245	0.296

P values were calculated using the roc test by Delong method. * statistical significant ($p < 0.05$).

Table S5. list of excluded features affected by inter-observer segmentation variability.

O.shape_Elongation	W.HLL_glszm_LowGrayLevelZoneEmphasis
O.firstorder_10Percentile	W.HLL_glszm_SmallAreaLowGrayLevelEmphasis
O.firstorder_Kurtosis	W.HLL_gldm_ LargeDependenceLowGrayLevelEmphasis
O.firstorder_Minimum	W.HLL_gldm_LowGrayLevelEmphasis
O.glcm_Correlation	W.HLL_gldm_ SmallDependenceLowGrayLevelEmphasis
O.glrlm_LongRunLowGrayLevelEmphasis	W.HLH_firstorder_Mean
O.glrlm_LowGrayLevelRunEmphasis	W.HLH_firstorder_Median
O.glrlm_ShortRunLowGrayLevelEmphasis	W.HLH_firstorder_RootMeanSquared
O.glszm_LowGrayLevelZoneEmphasis	W.HLH_firstorder_Skewness
O.glszm_SmallAreaLowGrayLevelEmphasis	W.HLH_glcm_ClusterShade
O.gldm_LargeDependenceHighGrayLevelEmphasis	W.HLH_glszm_SmallAreaLowGrayLevelEmphasis
O.gldm_LowGrayLevelEmphasis	W.HLH_gldm_ SmallDependenceLowGrayLevelEmphasis
O.gldm_SmallDependenceLowGrayLevelEmphasis	W.HLH_ngtdm_Contrast
O.ngtdm_Strength	W.HLH_ngtdm_Strength
W.LLH_firstorder_Mean	W.HHL_firstorder_Mean
W.LLH_firstorder_Median	W.HHL_firstorder_Median
W.LLH_firstorder_RootMeanSquared	W.HHL_firstorder_RootMeanSquared
W.LLH_firstorder_Skewness	W.HHL_firstorder_Skewness
W.LLH_glcm_ClusterShade	W.HHL_glcm_ClusterShade
W.LLH_ngtdm_Contrast	W.HHL_glrlm_LowGrayLevelRunEmphasis
W.LLH_ngtdm_Strength	W.HHL_glrlm_ShortRunLowGrayLevelEmphasis
W.LHL_firstorder_Mean	W.HHL_glszm_SizeZoneNonUniformityNormalized
W.LHL_firstorder_Median	W.HHL_glszm_SmallAreaEmphasis
W.LHL_firstorder_RootMeanSquared	W.HHL_gldm_LowGrayLevelEmphasis
W.LHH_firstorder_Mean	W.HHH_firstorder_Kurtosis
W.LHH_firstorder_Median	W.HHH_firstorder_Mean
W.LHH_firstorder_RootMeanSquared	W.HHH_firstorder_Median

W.LHH_firstorder_Skewness	W.HHH_firstorder_RootMeanSquared
W.LHH_glcm_ClusterShade	W.HHH_firstorder_Skewness
W.LHH_glcm_Imc1	W.HHH_glcm_ClusterShade
W.LHH_glrlm_LongRunLowGrayLevelEmphasis	W.HHH_glcm_Idmn
W.LHH_glrlm_LowGrayLevelRunEmphasis	W.HHH_glrlm_ShortRunLowGrayLevelEmphasis
W.LHH_glrlm_ShortRunLowGrayLevelEmphasis	W.HHH_glszm_LowGrayLevelZoneEmphasis
W.LHH_gldm_LowGrayLevelEmphasis	W.HHH_glszm_SizeZoneNonUniformityNormalized
W.LHH_gldm_ SmallDependenceLowGrayLevelEmphasis	W.HHH_glszm_SmallAreaEmphasis
W.LHH_ngtdm_Strength	W.HHH_glszm_SmallAreaLowGrayLevelEmphasis
W.HLL_firstorder_Skewness	W.HHH_gldm_ SmallDependenceLowGrayLevelEmphasis
W.HLL_glcm_ClusterShade	W.HHH_ngtdm_Strength
W.HLL_glcm_Correlation	W.LLL_firstorder_10Percentile
W.HLL_glcm_Idmn	W.LLL_firstorder_Kurtosis
W.HLL_glrlm_LongRunLowGrayLevelEmphasis	W.LLL_firstorder_Minimum
W.HLL_glrlm_LowGrayLevelRunEmphasis	W.LLL_glcm_Correlation
W.HLL_glrlm_ShortRunLowGrayLevelEmphasis	W.LLL_gldm_ LargeDependenceLowGrayLevelEmphasis
W.LLL_ngtdm_Strength	

Table S6. Ranked AUC values of the 100 radiomics, clinical and combined trainings models for the three strategies.

Radiomics			Clinical			Combined		
Strategy 1	**Strategy 2**	**Strategy 3**	**Strategy 1**	**Strategy 2**	**Strategy 3**	**Strategy 1**	**Strategy 2**	**Strategy 3**
0,55	0,46	0,49	0,72	0,68	0,74	0,74	0,59	0,73
0,56	0,49	0,52	0,73	0,70	0,75	0,75	0,72	0,73
0,62	0,53	0,52	0,74	0,70	0,75	0,76	0,73	0,76
0,63	0,53	0,53	0,74	0,70	0,75	0,77	0,73	0,77
0,64	0,55	0,54	0,74	0,71	0,75	0,77	0,75	0,77
0,66	0,55	0,55	0,75	0,71	0,75	0,78	0,75	0,78
0,66	0,56	0,55	0,76	0,71	0,76	0,78	0,75	0,78
0,67	0,57	0,57	0,76	0,72	0,76	0,79	0,76	0,78
0,68	0,58	0,58	0,76	0,72	0,76	0,79	0,76	0,79
0,68	0,59	0,58	0,76	0,72	0,76	0,79	0,77	0,79
0,68	0,59	0,60	0,76	0,73	0,77	0,80	0,77	0,79
0,69	0,60	0,60	0,76	0,73	0,77	0,80	0,77	0,79
0,69	0,60	0,61	0,77	0,73	0,77	0,80	0,77	0,79

Radiomics			Clinical			Combined		
Strategy 1	Strategy 2	Strategy 3	Strategy 1	Strategy 2	Strategy 3	Strategy 1	Strategy 2	Strategy 3
0,69	0,60	0,62	0,77	0,73	0,77	0,80	0,77	0,79
0,69	0,61	0,62	0,77	0,74	0,78	0,81	0,78	0,79
0,69	0,61	0,62	0,77	0,74	0,78	0,81	0,78	0,79
0,70	0,61	0,63	0,77	0,74	0,78	0,81	0,78	0,79
0,70	0,63	0,63	0,77	0,74	0,78	0,81	0,78	0,79
0,70	0,63	0,63	0,77	0,74	0,78	0,82	0,79	0,79
0,70	0,64	0,63	0,77	0,74	0,78	0,82	0,79	0,79
0,71	0,64	0,64	0,77	0,74	0,78	0,82	0,79	0,79
0,71	0,64	0,64	0,77	0,74	0,79	0,82	0,79	0,79
0,72	0,64	0,64	0,77	0,74	0,79	0,82	0,79	0,80
0,72	0,64	0,64	0,77	0,75	0,79	0,82	0,79	0,80
0,72	0,64	0,64	0,77	0,75	0,79	0,82	0,80	0,80
0,72	0,64	0,64	0,78	0,75	0,79	0,82	0,80	0,80
0,72	0,65	0,65	0,78	0,75	0,79	0,82	0,80	0,80
0,73	0,65	0,65	0,78	0,75	0,79	0,82	0,80	0,80
0,73	0,65	0,65	0,78	0,75	0,79	0,82	0,80	0,80
0,73	0,65	0,66	0,78	0,76	0,80	0,82	0,80	0,80
0,73	0,65	0,66	0,78	0,76	0,80	0,82	0,80	0,81
0,73	0,65	0,66	0,78	0,76	0,80	0,82	0,80	0,81
0,73	0,66	0,66	0,78	0,76	0,80	0,82	0,80	0,81
0,74	0,66	0,66	0,78	0,77	0,80	0,82	0,81	0,81
0,74	0,67	0,66	0,78	0,77	0,80	0,82	0,81	0,81
0,74	0,67	0,66	0,79	0,77	0,80	0,82	0,81	0,81
0,74	0,67	0,67	0,79	0,77	0,80	0,82	0,81	0,81
0,74	0,67	0,67	0,79	0,77	0,80	0,82	0,81	0,81
0,74	0,68	0,67	0,79	0,77	0,80	0,82	0,81	0,81
0,74	0,68	0,67	0,79	0,77	0,80	0,83	0,81	0,81
0,74	0,68	0,67	0,79	0,77	0,80	0,83	0,81	0,81
0,75	0,69	0,67	0,79	0,78	0,80	0,83	0,81	0,81
0,75	0,69	0,67	0,79	0,78	0,80	0,83	0,82	0,81
0,75	0,69	0,68	0,79	0,78	0,80	0,83	0,82	0,82
0,75	0,69	0,68	0,79	0,78	0,80	0,83	0,82	0,82
0,75	0,69	0,68	0,79	0,78	0,80	0,83	0,82	0,82
0,75	0,69	0,68	0,79	0,78	0,80	0,83	0,82	0,82
0,75	0,69	0,68	0,79	0,78	0,81	0,83	0,82	0,82
0,75	0,70	0,68	0,79	0,78	0,81	0,83	0,82	0,82
0,76	0,70	0,68	0,79	0,78	0,81	0,83	0,82	0,82

Radiomics			Clinical			Combined		
Strategy 1	Strategy 2	Strategy 3	Strategy 1	Strategy 2	Strategy 3	Strategy 1	Strategy 2	Strategy 3
0,76	0,70	0,68	0,79	0,79	0,81	0,83	0,82	0,82
0,76	0,70	0,69	0,79	0,79	0,81	0,83	0,83	0,82
0,76	0,71	0,69	0,79	0,79	0,81	0,83	0,83	0,82
0,76	0,71	0,69	0,79	0,79	0,81	0,83	0,83	0,82
0,76	0,71	0,69	0,79	0,79	0,81	0,84	0,83	0,82
0,76	0,71	0,69	0,80	0,79	0,81	0,84	0,83	0,82
0,76	0,71	0,70	0,80	0,79	0,81	0,84	0,83	0,82
0,76	0,71	0,70	0,80	0,80	0,81	0,84	0,83	0,82
0,77	0,71	0,70	0,80	0,80	0,81	0,84	0,83	0,83
0,77	0,71	0,70	0,80	0,80	0,81	0,84	0,83	0,83
0,77	0,72	0,70	0,80	0,80	0,81	0,84	0,83	0,83
0,77	0,72	0,70	0,80	0,80	0,81	0,84	0,83	0,83
0,77	0,72	0,70	0,80	0,80	0,81	0,84	0,84	0,83
0,77	0,72	0,71	0,80	0,80	0,82	0,84	0,84	0,83
0,77	0,72	0,71	0,80	0,80	0,82	0,84	0,84	0,83
0,78	0,72	0,71	0,80	0,80	0,82	0,85	0,84	0,83
0,78	0,72	0,71	0,80	0,80	0,82	0,85	0,84	0,83
0,78	0,73	0,71	0,81	0,80	0,82	0,85	0,84	0,83
0,78	0,73	0,71	0,81	0,80	0,82	0,85	0,84	0,83
0,78	0,73	0,72	0,81	0,80	0,82	0,85	0,84	0,83
0,78	0,73	0,72	0,81	0,81	0,82	0,85	0,84	0,83
0,78	0,73	0,72	0,81	0,81	0,82	0,85	0,84	0,83
0,79	0,73	0,72	0,81	0,81	0,82	0,85	0,85	0,83
0,79	0,73	0,72	0,82	0,81	0,82	0,85	0,85	0,83
0,79	0,73	0,73	0,82	0,81	0,82	0,85	0,85	0,84
0,79	0,73	0,73	0,82	0,81	0,82	0,85	0,85	0,84
0,79	0,74	0,73	0,82	0,81	0,83	0,86	0,85	0,84
0,79	0,74	0,73	0,82	0,81	0,83	0,86	0,85	0,84
0,79	0,74	0,73	0,82	0,81	0,83	0,86	0,85	0,84
0,79	0,75	0,73	0,82	0,81	0,83	0,86	0,85	0,84
0,79	0,75	0,73	0,82	0,82	0,83	0,86	0,85	0,84
0,79	0,75	0,73	0,83	0,82	0,83	0,86	0,85	0,84
0,79	0,75	0,73	0,83	0,82	0,83	0,86	0,85	0,84
0,80	0,75	0,74	0,83	0,82	0,83	0,86	0,85	0,84
0,80	0,75	0,74	0,83	0,82	0,84	0,86	0,85	0,85
0,80	0,75	0,75	0,83	0,82	0,84	0,86	0,86	0,85
0,80	0,76	0,75	0,83	0,82	0,84	0,86	0,86	0,85

Radiomics			Clinical			Combined		
Strategy 1	Strategy 2	Strategy 3	Strategy 1	Strategy 2	Strategy 3	Strategy 1	Strategy 2	Strategy 3
0,80	0,76	0,75	0,84	0,83	0,84	0,87	0,86	0,85
0,80	0,77	0,75	0,84	0,84	0,84	0,87	0,86	0,85
0,81	0,77	0,75	0,84	0,84	0,85	0,87	0,86	0,85
0,81	0,77	0,75	0,85	0,84	0,85	0,87	0,86	0,86
0,81	0,77	0,75	0,85	0,84	0,85	0,87	0,86	0,86
0,81	0,77	0,75	0,85	0,84	0,85	0,87	0,87	0,86
0,81	0,77	0,76	0,85	0,84	0,85	0,87	0,87	0,86
0,82	0,78	0,76	0,85	0,85	0,86	0,87	0,87	0,86
0,82	0,78	0,76	0,85	0,86	0,86	0,87	0,87	0,86
0,82	0,78	0,77	0,86	0,86	0,86	0,88	0,88	0,87
0,83	0,78	0,77	0,86	0,87	0,87	0,88	0,88	0,87
0,85	0,79	0,77	0,86	0,87	0,87	0,91	0,88	0,87
0,86	0,81	0,77	0,87	0,87	0,88	0,91	0,88	0,88

Table S7. Radiomics Quality Score.

Criteria	Points
Image protocol quality	+ 1
Multiple segmentations	+ 1
Phantom study on all scanners	+ 0
Imaging at multiple time points	+ 0
Feature reduction or adjustment for multiple testing	+ 3
Multivariate analysis with non radiomics features	+ 1
Detect and discuss biological correlates	+ 0
Cut-off analyses	+ 0
Discrimination statistics	+ 2
Calibration statistics	+ 0
Prospective study registered in a trial database	+ 0
Validation	+ 3
Comparison to 'gold standard'	+ 2
Potential clinical utility	+ 2
Cost-effectiveness analysis	+ 0
Open science and data	+ 0
Total	**15**

A total of 36 points can be achieved, with higher scores indicating higher research quality

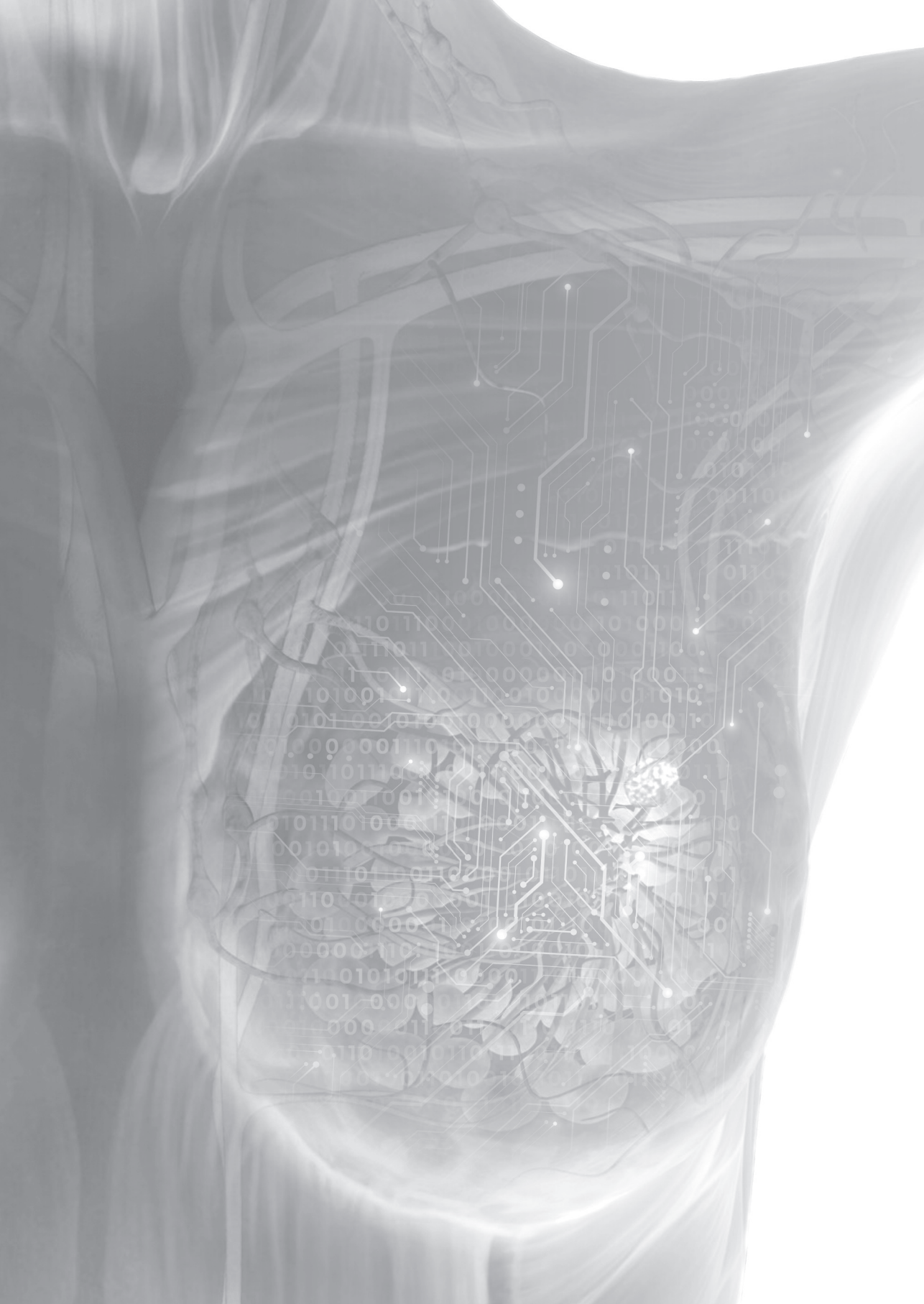
References

1. Vugts G, Maaskant-Braat AJ, Nieuwenhuijzen GA, Roumen RM, Luiten EJ, Voogd AC. Patterns of Care in the Administration of Neo-adjuvant Chemotherapy for Breast Cancer. A Population-Based Study. *The breast journal.* 2016;22(3):316-321.
2. Murphy BL, Day CN, Hoskin TL, Habermann EB, Boughey JC. Neoadjuvant Chemotherapy Use in Breast Cancer is Greatest in Excellent Responders: Triple-Negative and HER2+ Subtypes. *Annals of surgical oncology.* 2018;25(8):2241-2248.
3. Spronk PER, de Ligt KM, van Bommel ACM, et al. Current decisions on neoadjuvant chemotherapy for early breast cancer: Experts' experiences in the Netherlands. *Patient Educ Couns.* 2018.
4. Loibl S, Denkert C, von Minckwitz G. Neoadjuvant treatment of breast cancer--Clinical and research perspective. *Breast (Edinburgh, Scotland).* 2015;24 Suppl 2:S73-77.
5. Prevos R, Smidt ML, Tjan-Heijnen VC, et al. Pre-treatment differences and early response monitoring of neoadjuvant chemotherapy in breast cancer patients using magnetic resonance imaging: a systematic review. *Eur Radiol.* 2012;22(12):2607-2616.
6. Lobbes MB, Prevos R, Smidt M, et al. The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: a systematic review. *Insights Imaging.* 2013;4(2):163-175.
7. Weber JJ, Jochelson MS, Eaton A, et al. MRI and Prediction of Pathologic Complete Response in the Breast and Axilla after Neoadjuvant Chemotherapy for Breast Cancer: MRI and Pathologic Complete Response. *J Am Coll Surg.* 2017.
8. Bouzon A, Acea B, Soler R, et al. Diagnostic accuracy of MRI to evaluate tumour response and residual tumour size after neoadjuvant chemotherapy in breast cancer patients. *Radiol Oncol.* 2016;50(1):73-79.
9. van der Noordaa MEM, van Duijnhoven FH, Loo CE, et al. Identifying pathologic complete response of the breast after neoadjuvant systemic therapy with ultrasound guided biopsy to eventually omit surgery: Study design and feasibility of the MICRA trial (Minimally Invasive Complete Response Assessment). *Breast (Edinburgh, Scotland).* 2018;40:76-81.
10. Heil J, Sinn P, Richter H, et al. RESPONDER - diagnosis of pathological complete response by vacuum-assisted biopsy after neoadjuvant chemotherapy in breast Cancer - a multicenter, confirmative, one-armed, intra-individually-controlled, open, diagnostic trial. *BMC Cancer.* 2018;18(1):851.
11. van Loevezijn AA, van der Noordaa MEM, van Werkhoven ED, et al. Minimally Invasive Complete Response Assessment of the Breast After Neoadjuvant Systemic Therapy for Early Breast Cancer (MICRA trial): Interim Analysis of a Multicenter Observational Cohort Study. *Annals of surgical oncology.* 2020.
12. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017;77(21):e104-e107.
13. Ibrahim A, Primakov S, Beuque M, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods.* 2020.
14. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749-762.

15. Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol.* 2017;90(1070):20160665.
16. Refaee T, Wu G, Ibrahim A, et al. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration.* 2020;99(2):99-107.
17. Ahmed A, Gibbs P, Pickles M, Turnbull Lr. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *J Magn Reson Imaging.* 2013;38(1):89-101.
18. Parikh J, Selmi M, Charles-Edwards G, et al. Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy. *Radiology.* 2014;272(1):100-112.
19. Teruel JR, Heldahl MG, Goa PE, et al. Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *NMR Biomed.* 2014;27(8):887-896.
20. Chamming's F, Ueno Y, Ferre R, et al. Features from Computerized Texture Analysis of Breast Cancers at Pretreatment MR Imaging Are Associated with Response to Neoadjuvant Chemotherapy. *Radiology.* 2018;286(2):412-420.
21. Yoon HJ, Kim Y, Chung J, Kim BSr. Predicting neo-adjuvant chemotherapy response and progression-free survival of locally advanced breast cancer using textural features of intratumoral heterogeneity on F-18 FDG PET/CT and diffusion-weighted MR imaging. *The breast journal.* 2018.
22. Baessler B, Weiss K, Pinto Dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest Radiol.* 2019;54(4):221-228.
23. Mackin D, Fave X, Zhang L, et al. Measuring CT scanner variability of radiomics features. *Investigative radiology.* 2015;50(11):757.
24. Rai R, Holloway LC, Brink C, et al. Multicenter evaluation of MRI-based radiomic features: A phantom study. *Med Phys.* 2020.
25. Ibrahim A, Refaee T, Primakov S, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers.* 2021;13(8).
26. Ibrahim A, Refaee T, Leijenaar RTH, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS One.* 2021;16(5):e0251147.
27. Dreher C, Kuder TA, Konig F, et al. Radiomics in diffusion data: a test-retest, inter- and intra-reader DWI phantom study. *Clin Radiol.* 2020;75(10):798 e713-798 e722.
28. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep.* 2019;9(1):4800.
29. Shur J, Blackledge M, D'Arcy J, et al. MRI texture feature repeatability and image acquisition factor robustness, a phantom study and in silico study. *Eur Radiol Exp.* 2021;5(1):2.
30. Ogston KN, Miller ID, Payne S, et al. A new histological grading system to assess response of breast cancers to primary chemotherapy: prognostic significance and survival. *Breast (Edinburgh, Scotland).* 2003;12(5):320-327.
31. Pinder SE, Provenzano E, Earl H, Ellis IO. Laboratory handling and histology reporting of breast specimens from patients who have received neoadjuvant chemotherapy. *Histopathology.* 2007;50(4):409-417.

32. Fan M, Wu G, Cheng H, Zhang J, Shao G, Li Lr. Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients. *European journal of radiology.* 2017;94:140-147.
33. Braman N, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res Journal Translated Name Breast Cancer Research.* 2017;19(1):no pagination.
34. Granzier RWY, Verbakel NMH, Ibrahim A, et al. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. *Sci Rep.* 2020;10(1):14163.
35. Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging.* 2000;19(2):143-150.
36. Hoebel KV, Patel JB, Beers AL, et al. Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma. *Radiology: Artificial Intelligence.* 2021;3(1).
37. Ligero M, Jordi-Ollero O, Bernatowicz K, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur Radiol.* 2020.
38. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *J Appl Clin Med Phys.* 2020;21(1):179-190.
39. Sun X, Shi L, Luo Y, et al. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomed Eng Online.* 2015;14:73.
40. Senthilkumaran N, Thimmiaraja J. Histogram Equalization for Image Enhancement Using MRI Brain Images. *2014 World Congress on Computing and Communication Technologies.* 2014:pp. 80-83.
41. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS One.* 2019;14(3):e0213459.
42. Kursa MB, Jankowski A, Rudnicki WR. Boruta - A System for Feature Selection. *Fund Inform.* 2010;101(4):271-286.
43. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw.* 2010;36(11):1-13.
44. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci.* 2003;43(6):1947-1958.
45. Racine JS. RStudio: a platform-independent IDE for R and Sweave. *Journal of Applied Econometrics.* 2012;27(1):167-172.
46. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12:77.
47. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
48. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* 2015;13:1.

49. Bianchini L, Botta F, Origgi D, et al. PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis. *Phys Med.* 2020;71:71-81.
50. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci Rep.* 2019;9(1):9441.
51. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2019;135:107-114.
52. Scalco E, Belfatto A, Mastropietro A, et al. T2w-MRI signal normalization affects radiomics features reproducibility. *Med Phys.* 2020;47(4):1680-1691.
53. Choudhery S, Gomez-Cardona D, Favazza CP, et al. MRI Radiomics for Assessment of Molecular Subtype, Pathological Complete Response, and Residual Cancer Burden in Breast Cancer Patients Treated With Neoadjuvant Chemotherapy. *Acad Radiol.* 2020.
54. Henderson S, Purdie C, Michie C, et al. Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer. *Eur Radiol.* 2017;27(11):4602-4611.
55. Cain EH, Saha A, Harowicz MR, Marks JR, Marcom PK, Mazurowski MA. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast cancer research and treatment.* 2018.
56. Xiong Q, Zhou X, Liu Z, et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy. *Clin Transl Oncol.* 2019.
57. Liu Z, Li Z, Qu J, et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res.* 2019.



CHAPTER 5

Dedicated axillary MRI-based radiomics analysis for the prediction of axillary lymph node metastasis in breast cancer

Sanaz Samiei*, Renée W. Y. Granzier*, Abdalla Ibrahim, Sergey Primakov, Marc B. I. Lobbes, Regina G. H. Beets-Tan, Thiemo J. A. van Nijnatten, Sanne M. E. Engelen, Henry C. Woodruff, and Marjolein L. Smidt.

*Shared first authorship

Cancers. 2021 Feb;13(4):757

Abstract

Radiomics features may contribute to increased diagnostic performance of MRI in the prediction of axillary lymph node metastasis. The objective of the study was to predict preoperative axillary lymph node metastasis in breast cancer using clinical models and radiomics models based on T2-weighted (T2W) dedicated axillary MRI features with node-by-node analysis. From August 2012 until October 2014, all women who had undergone dedicated axillary 3.0T T2W MRI, followed by axillary surgery, were retrospectively identified, and available clinical data were collected. All axillary lymph nodes were manually delineated on the T2W MR images, and quantitative radiomics features were extracted from the delineated regions. Data were partitioned patient-wise to train 100 models using different splits for the training and validation cohorts to account for multiple lymph nodes per patient and class imbalance. Features were selected in the training cohorts using recursive feature elimination with repeated 5-fold cross-validation, followed by the development of random forest models. The performance of the models was assessed using the area under the curve (AUC). A total of 75 women (median age, 61 years; interquartile range, 51–68 years) with 511 axillary lymph nodes were included. On final pathology, 36 (7%) of the lymph nodes had metastasis. A total of 105 original radiomics features were extracted from the T2W MR images. Each cohort split resulted in a different number of lymph nodes in the training cohorts and a different set of selected features. Performance of the 100 clinical and radiomics models showed a wide range of AUC values between 0.41–0.74 and 0.48–0.89 in the training cohorts, respectively, and between 0.30–0.98 and 0.37–0.99 in the validation cohorts, respectively. With these results, it was not possible to obtain a final prediction model. Clinical characteristics and dedicated axillary MRI-based radiomics with node-by-node analysis did not contribute to the prediction of axillary lymph node metastasis in breast cancer based on data where variations in acquisition and reconstruction parameters were not addressed.

Introduction

In breast cancer patients, the axillary lymph node status provides essential prognostic information about the locoregional recurrence and overall survival rate[1-4]. The five-year survival rate decreases from 99% to 85% with the presence of lymph node metastasis in the axilla[5]. The presence of axillary lymph node metastasis determines the extent of the surgical treatment plan, the potential need for (neo) adjuvant systemic therapy, and the possible indication for postmastectomy radiation therapy with regard to immediate breast reconstruction[6,7].

In the preoperative setting, imaging for axillary lymph node assessment is recommended in the clinical workup of invasive breast cancer patients[6]. For the evaluation of tumor extent in the breast or following neoadjuvant treatment, breast magnetic resonance imaging (MRI) is often performed, which includes the axilla in the field of view[8]. However, when using dedicated breast coils, the field of view of the axillary region can be limited[9]. Therefore, dedicated MR coils for visualization and assessment of the axillary region have been investigated[10-12]. Dedicated unenhanced T2-weighted (T2W) axillary MRI showed good diagnostic performance based on node-by-node analysis but remained insufficient to accurately exclude axillary lymph node metastasis[12].

Although preoperative imaging may be performed to guide the axillary management of patients, no current imaging modality with optimal diagnostic performance can replace the surgical axillary staging procedure. In the era of artificial intelligence, current developments in radiology focus on the improvement of decision support systems to maximize the potential role of noninvasive imaging modalities. Radiomics, the application of machine learning to medical imaging, is a rapidly evolving field that enables high-throughput quantitative data extraction from standard medical images in an automated fashion and subsequent data analysis, possibly combined with patient and tumor characteristics, improving the accuracy of diagnostic, predictive, and prognostic models[13,14]. The evaluation of the usefulness of radiomics based on mammography, ultrasound, and breast MRI has been explored, showing potential in axillary lymph node metastasis prediction[15-19]. However, this research focused on the prediction of axillary lymph node metastasis from the delineated breast tumor as the region of interest (ROI), and not from the lymph nodes themselves.

Accurate preoperative prediction of axillary lymph node metastasis in breast cancer patients can assist in clinical decision-making regarding the type of treatment. Radiomics features extracted from axillary lymph nodes may contribute to increased diagnostic performance of MRI in the prediction of metastasis. To our knowledge, no previous study has reported on node-by-node matching of axillary lymph nodes

with pathological findings in breast cancer patients in the field of radiomics. The purpose of this study was to predict preoperative axillary lymph node metastasis in breast cancer patients using clinical models and radiomics models based on unenhanced T2W dedicated axillary MRI features with node-by-node analysis.

Materials and methods

Patient population

Consecutive women with histopathologically proven breast cancer, who had undergone dedicated axillary MRI between August 2012 and October 2014, followed by sentinel lymph node biopsy (SNLB) or axillary lymph node dissection (ALND), were considered for inclusion. Patients were excluded if they had undergone neoadjuvant systemic therapy before axillary surgery and in the case of ductal carcinoma in situ only. This study was approved by the local medical ethics committee, and the requirement of written informed consent was waived due to the retrospective study design. Fifty of the dedicated axillary T2W and diffusion-weighted MR images were earlier described by Schipper et al. for axillary lymph node staging, and 90 of the dedicated axillary T2W and gadofosveset-enhanced MR images were earlier described by van Nijnatten et al. for axillary lymph node staging[12,20].

Clinical and pathological characteristics

Clinical and pathological data were derived from the patients' medical records: age, clinical TNM stage, pathological TNM stage, tumor histology, tumor grade, breast cancer subtype, and type of axillary surgery. Lymph nodes with isolated tumor cells (£0.2 mm) and micrometastases (>0.2–£2.0 mm) were considered negative, while those with macrometastases (>2.0 mm) were considered positive.

MRI acquisition

The dedicated axillary MR images were performed using a 32-channel cardiac coil on a 3.0 Tesla scanner (Achieva, Philips Healthcare, Best, the Netherlands). During the MRI examination, the patient was positioned in a supine position with the ipsilateral arm elevated. The anatomical confines of the dedicated axillary MR images were between the humeral head and the inferior border of the scapula. The MRI protocol included an unenhanced three-dimensional T2W turbo spin-echo sequence without fat suppression (pixel size, 1.25 × 1.25 mm; repetition time, 2000 ms; echo time between 150-202 ms; echo train length, 52 or 66; flip angle, 90°; acquisition slice thickness, 2.5 mm; reconstruction slice thickness, 1.25 mm), a contrast-enhanced T1-weighted sequence, and a diffusion-weighted imaging sequence with fat suppression.

MRI lymph node delineation

All axillary lymph nodes of each dedicated axillary T2W MR image were manually delineated in three dimensions using MIM software (version 6.9.4, MIM Software Inc., Cleveland, OH, USA) by a medical researcher (S.S.) with three years of experience in axillary lymph node imaging validated by a dedicated breast radiologist (M.L.) with eleven years of experience (Figure 1). No clinical information and pathology results were available during delineation and validation. The delineated lymph nodes were subsequently matched with their histopathological findings (node-by-node matching). Reliable node-by-node matching was obtained using single-photon emission computed tomography-X-ray computed tomography (SPECT-CT) in patients undergoing SLNB, and an anatomical map was used for patients undergoing ALND. The exact procedure of the node-by-node matching was previously described by Schipper et al.[21].

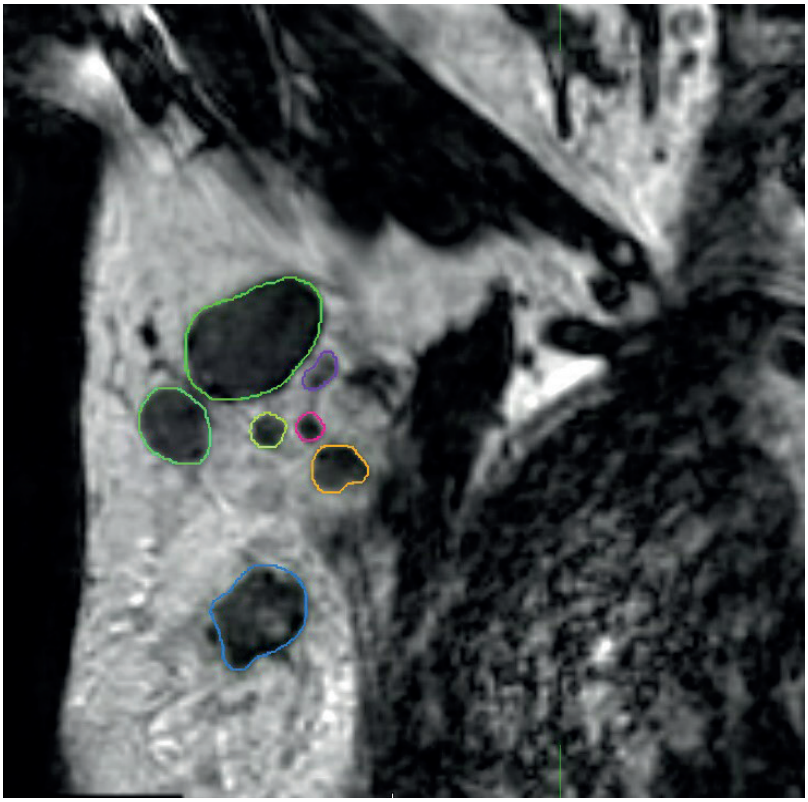


Figure 1. Coronal T2-weighted dedicated axillary MR image of a 55-year old woman with invasive breast cancer, who was treated with mastectomy and axillary lymph node dissection (pT1N2). The MR image demonstrates an example of delineations of lymph nodes in the right axilla on the MIM software.

MRI preprocessing and feature extraction

Image preprocessing of the T2W images was performed after delineation. Bias field correction was applied to every T2W MR image using MIM software to correct for non-uniform grayscale intensities caused by field inhomogeneities. To ensure better comparability of voxel intensities, additional image normalization and discretization was performed by the open-source Pyradiomics software (version 2.2.0) prior to feature extraction[22]. For discretization, grayscale values were aggregated with a fixed bin width of 10, which ensured the recommended amount of bins between 30–130[22]. Resampling was not required, as all images consisted of isotropic voxels of equal size 1.25 mm^3. Quantitative radiomics features were extracted from the delineated regions using the Pyradiomics software. The extracted features can be subdivided into the following classes: first-order statistics, three-dimensional shape-based, gray level co-occurrence matrix, gray level run length matrix, gray level size zone matrix, neighboring gray-tone difference matrix, and gray level dependence matrix.

Radiomics feature selection and model development

Taking into account the small skewed dataset and the unavailability of an external validation dataset, the data were randomly divided into training and validation cohort 100 times using two different strategies to create a more balanced training cohort. In the first strategy, 85% (12 out of 14) of the node-positive (i.e., patients with axillary lymph node metastasis at final pathology) breast cancer patients were selected in the training cohort, and all remaining node-positive and node-negative (i.e., patients without axillary lymph node metastasis at final pathology) patients in the validation cohort, considering each axillary lymph node as an individual data point when training the model. In the second strategy, only the lymph nodes of patients with node-positive breast cancer were considered as individual data points when training and validating the model. To maintain the original class imbalance of the node-positive patients, 10 patients were selected in the training cohort. For both strategies, additional models were developed using a random undersampled balanced training cohort. All lymph nodes of one patient were always included in either the training cohort or the validation cohort, and therefore each split caused a varying number of positive lymph nodes in each cohort. Feature selection started with the removal of near-zero variance features followed by the removal of highly correlated features using the Pearson pairwise correlation greater than 0.95. Subsequently, recursive feature elimination with bagged trees was applied with repeated 5-fold cross-validation to select a maximum number of features in the training cohort. The number of features was chosen at the point when the addition of more features did not increase the diagnostic performance of the models. Random forest binary classification models were trained, using optimized random forest parameters (number of

trees and features per node) for the training cohort, selecting the optimal number of features for each generated model. In addition, a separate set of models was generated using the same pipeline but by adding an additional feature selection step at the very beginning. In this step, features robust to the variability of manual delineations of breast tumors on MRI by four observers were selected according to three different cut-off values (intraclass correlation coefficient of >0.75, >0.80, and >0.90)[23]. Figure 2 provides an overview of strategies 1 and 2 with the different developed models.

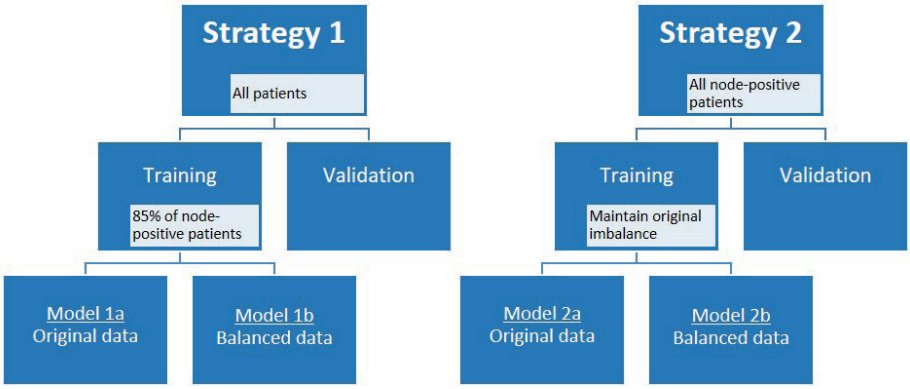


Figure 2. Model strategies.

Radiomics subanalysis

A separate set of models was generated using the first and second strategies as described earlier on a dataset where ROIs with less than 50 voxels were excluded[22]. On these models, only the additional feature selection step with different intraclass correlation coefficient cut-off values was not performed.

Clinical model development

Clinical models were trained based on clinical characteristics available before the axillary surgery. Random forest models with bagged tree function for the prediction of axillary lymph node metastasis were trained and validated using the same strategies as described above, except for the feature selection step, which was only the removal of highly correlated clinical characteristics. These clinical models were used to indicate the effect of known and unknown patient's biological covariates compared to a pure imaging-based model as well as to rank the importance of the clinical characteristics in this dataset using the Gini impurity method.

Statistical analyses and study evaluation

The statistical analyses, including dataset splitting and balancing, feature selection, model development, and performance evaluation, were performed in R (version 3.6.3; http//www.r-project.org) using R studio (version 1.2.1335, Vienna, Austria)[24]. The performance of all models was assessed using the area under the receiver operating characteristics curve (AUC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The Spearman correlation was used to calculate the correlation between the number of voxels per ROI and the corresponding pathological outcome. The radiomics workflow was evaluated using the radiomics quality score (RQS)[25]. This study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines[26].

Results

Patients characteristics

A total of ninety women were considered for inclusion, of whom twelve were excluded due to treatment with neoadjuvant systemic therapy before axillary surgery and three with ductal carcinoma in situ only. Seventy-five patients (median age, 61 years; interquartile range, 51–68 years) with 511 axillary lymph nodes were included. Patient, tumor, and treatment characteristics are summarized in Table 1. The median number of axillary lymph nodes per patient was six, with a range of 1–18. Fourteen of the included patients were node-positive at final pathology, with a total of 36 axillary lymph nodes with macrometastases and 58 axillary lymph nodes without metastasis. The remaining 61 patients had 417 axillary lymph nodes without metastasis. The median number of voxels per ROI for all delineated axillary lymph nodes was 100 (interquartile range, 44–236) and 310 (interquartile range, 130–1676) for all delineated axillary lymph nodes with metastasis. The Spearman correlation between the number of voxels per ROI and the corresponding pathological outcome was 0.22.

Radiomics feature extraction and model development

A total of 105 original radiomics features were extracted from the dedicated axillary T2W MR images. No near-zero variance features were detected. Pearson pairwise correlation removed 53 highly correlated features. The optimal subset of features was selected in the training cohort using recursive feature elimination with repeated 5-fold cross-validation with a maximum of 20 features. Figure 3 shows the distribution of the number of selected features from the 100 iterations for the two different strategies (lymph nodes from all patients versus only lymph nodes from node-positive patients as data points) for each model. Supplementary Material A includes a list of how often each feature was chosen in the 100 iterations for each model.

Table 1. Patient, tumor, and treatment characteristics.

Characteristic	Value
No. of patients	75
Age (years) (median; IQR)	61 (51–68)
Clinical tumor size (mm) (median, IQR)	19 (13–28)
Clinical tumor stage (%)	
T1	41 (54.7)
T2	32 (42.7)
T3	2 (2.6)
Clinical nodal stage (%)	
N0	68 (90.7)
N1	7 (9.3)
Tumor histology (%)	
Invasive ductal	55 (73.3)
Invasive lobular	11 (14.7)
Mixed invasive ductal & lobular	3 (4.0)
Other	6 (8.0)
Tumor grade (%)	
1	17 (22.7)
2	42 (56.0)
3	16 (21.3)
Breast cancer subtype (%)	
ER + HER2−	55 (73.3)
ER + HER2+	6 (9.0)
ER − HER2+	2 (2.7)
Triple-negative	11 (14.7)
Not determined	1 (1.3)
Axillary surgery (%)	
SLNB	8 (10.7)
ALND	67 (89.3)

Abbreviations: ER, Estrogen receptor; HER2, Human epidermal growth factor receptor 2; IQR, interquartile range; SLNB, Sentinel lymph node biopsy; ALND, Axillary lymph node dissection.

As each iteration resulted in a different set of selected features for each model in both strategies, it was not possible to obtain a final prediction model. The minimum and maximum area under the curve (AUC) values in the training cohorts were 0.59–0.80, 0.60–0.85, 0.48–0.84, and 0.55–0.89 for models 1a, 1b, 2a, and 2b, respectively. The median AUC values for all models in the training cohorts were between 0.72–0.73. All models showed a wider range of AUC values in the validation cohorts. The AUC value distribution for all models in the training and validation cohorts are presented in the violin plots in Figure 4. The minimum and maximum sensitivity in the training cohorts were 30–66%, 53–83%, 7–74%, and 48–82% for models 1a, 1b, 2a, and 2b, respectively. The median sensitivity for all models in the training cohorts was between 47–66%. All models showed lower median sensitivity in the validation cohorts. The minimum and maximum PPV in the training cohorts were 46–78%, 55–83%, 25–80%, and 52–90% for models 1a, 1b, 2a, and 2b, respectively. The median PPV for all models in the training cohorts were between 61–67%. All models showed a lower median PPV in the validation

cohorts. The diagnostic performance parameters of the radiomics models (100 iterations) are shown in Table 2.

The additional feature selection step with the cut-off values >0.75, >0.80, and >0.90 resulted in 44, 35, and 8 original features, respectively, available for recursive feature elimination with repeated 5-fold cross-validation. These results showed no differences compared to the results found without this additional feature selection step. The violin plots of the models developed after adding the additional feature selection step can be found in Figures S1–S3.

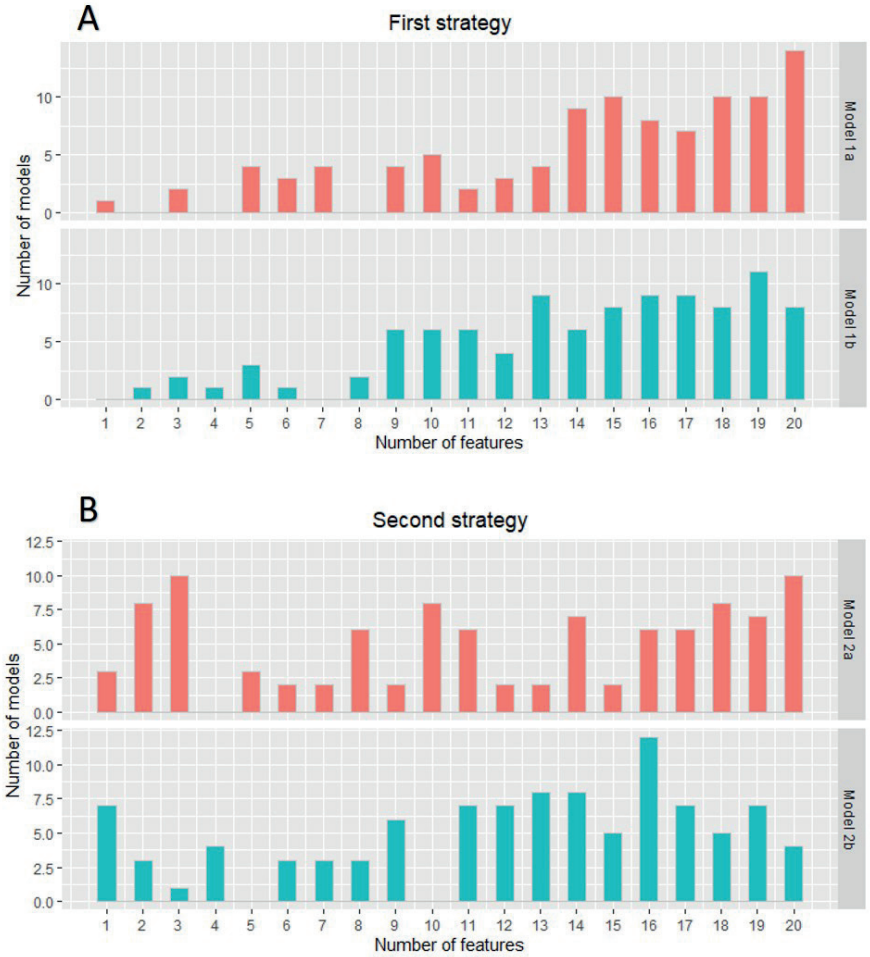


Figure 3. First (***A***) and second (***B***) strategy: distribution of the number of features in each developed model. The two different models in both strategies were all developed 100 times.

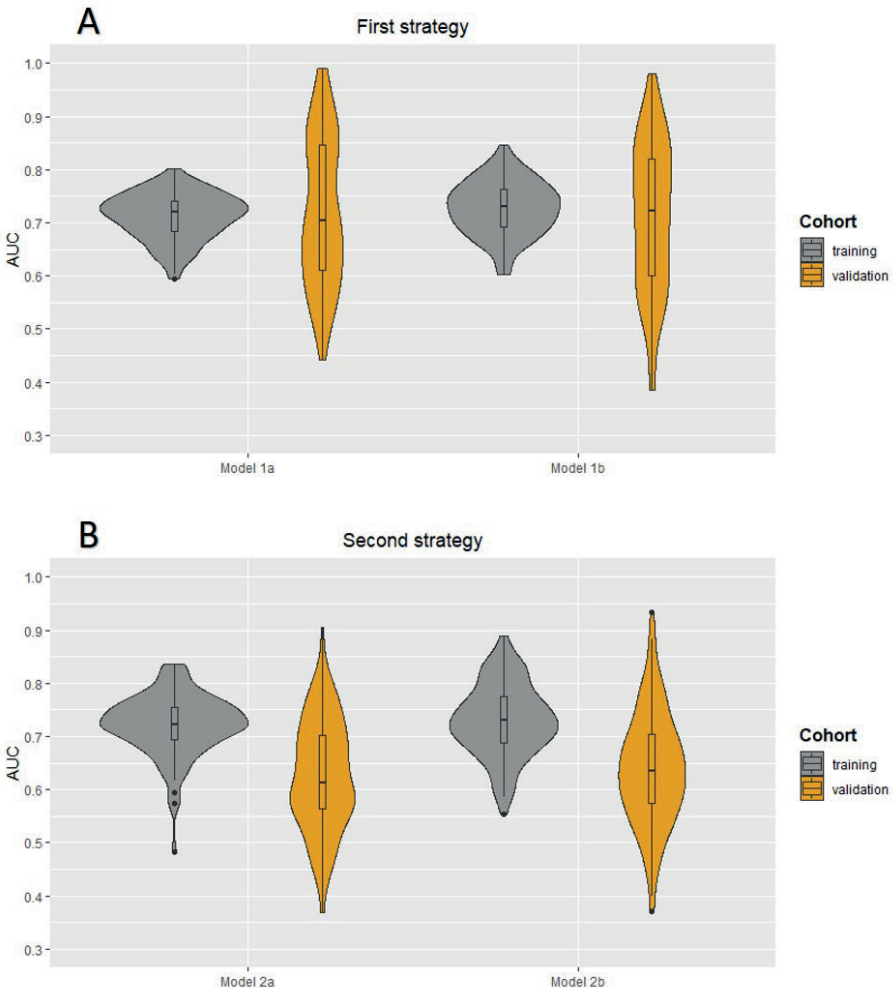


Figure 4. Violin plots for the radiomics models developed using the first (**A**) and second (**B**) strategy: AUC value distributions (100 iterations) for the four models (1a, 1b, 2a, and 2b) in both the training and validation cohort.

Table 2. The diagnostic performance of the radiomics models (100 iterations) for the first and second strategy.

Diagnostic parameters	Training				Validation				Training				Validation			
	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)
									First Strategy							
	Model 1a								Model 1b							
Minimum	30	71	46	62	0	78	0	98	53	50	55	72	0	57	0	98
Median	47	81	61	72	33	90	2	99	66	67	67	80	50	75	1	99
Maximum	66	91	78	79	100	97	22	100	83	85	83	88	100	88	10	100
									Second Strategy							
	Model 2a								Model 2b							
Minimum	7	58	25	54	0	33	0	22	48	46	52	68	0	0	0	0
Median	50	81	62	74	33	76	50	71	66	68	67	80	64	60	50	75
Maximum	74	93	80	83	82	100	100	88	82	92	90	89	100	100	100	100

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; sens, sensitivity; spec, specificity.
The additional feature selection step with the cut-off values >0.75, >0.80, and >0.90 resulted in 44, 35, and 8 original features, respectively, available for recursive feature elimination with repeated 5-fold cross-validation. These results showed no differences compared to the results found without this additional feature selection step. The violin plots of the models developed after adding the additional feature selection step can be found in Figures S1–S3.

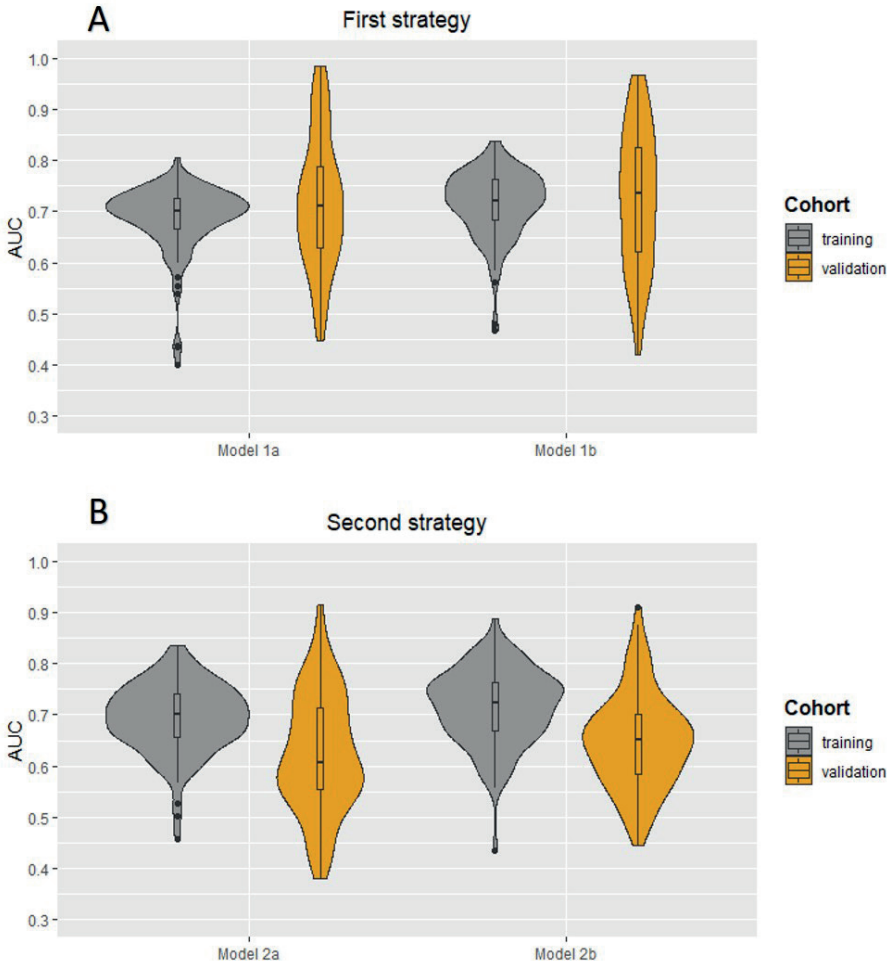


Figure S1. Violin plots for the radiomics models developed using the first (A) and second (B) strategy with additional feature se-lection step (ICC > 0.75): AUC value distributions (100 iterations) for the four models (1a, 1b, 2a and 2b) in both the training and validation cohort.

Radiomics subanalysis

After the exclusion of ROIs with less than 50 voxels, a total of 71 patients were included for analyses, with 371 axillary lymph nodes. Thirteen of these patients were node-positive, with a total of 31 axillary lymph nodes with metastasis and 34 axillary lymph nodes without metastases. The remaining 58 patients had 340 axillary lymph nodes without metastasis. Excluding small lymph nodes resulted in balanced training cohorts in models 1a and 2a, eliminating the need to perform

random undersampling (models 1b and 2b). The minimum and maximum AUC values of the balanced models 1a and 2a in the training and validation cohorts of this subanalysis were 0.53–0.82 and 0.41–0.83, respectively. Violin plots with the distribution of the AUC values and the diagnostic performance parameters of the subanalysis are provided in Table S1 and Figure S4.

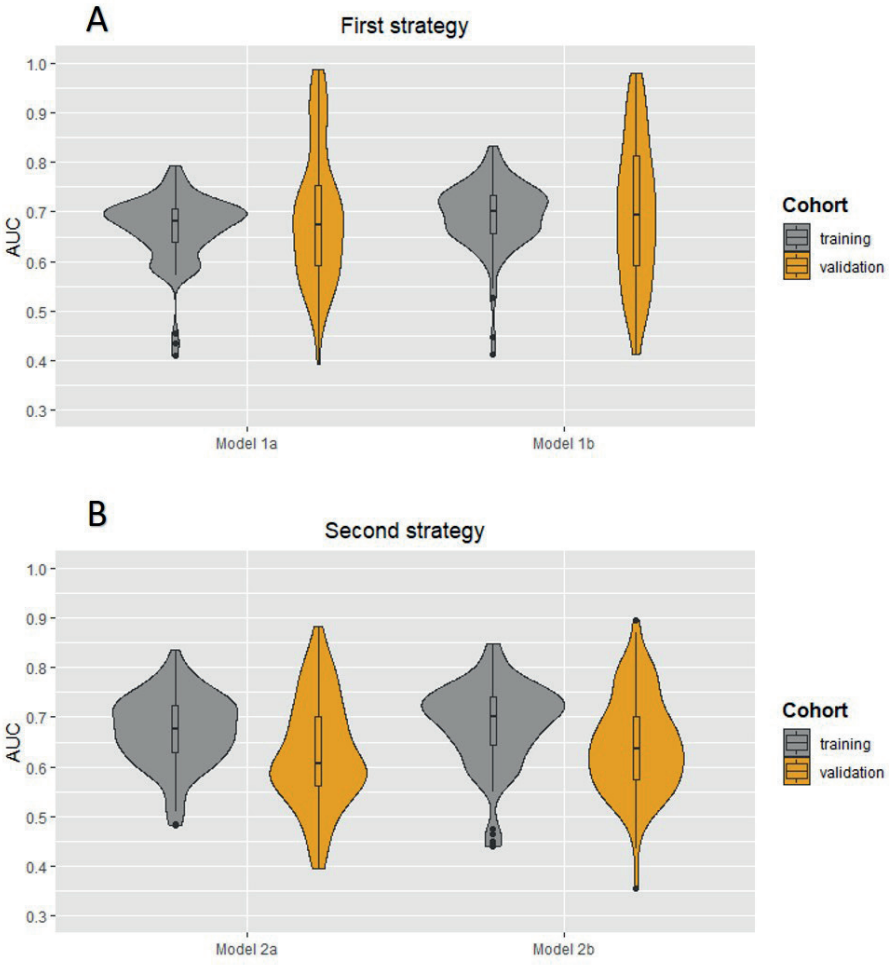


Figure S2. Violin plots for the radiomics models developed using the first (A) and second (B) strategy with additional feature se-lection step (ICC > 0.80): AUC value distributions (100 iterations) for the four models (1a, 1b, 2a and 2b) in both the training and validation cohort.

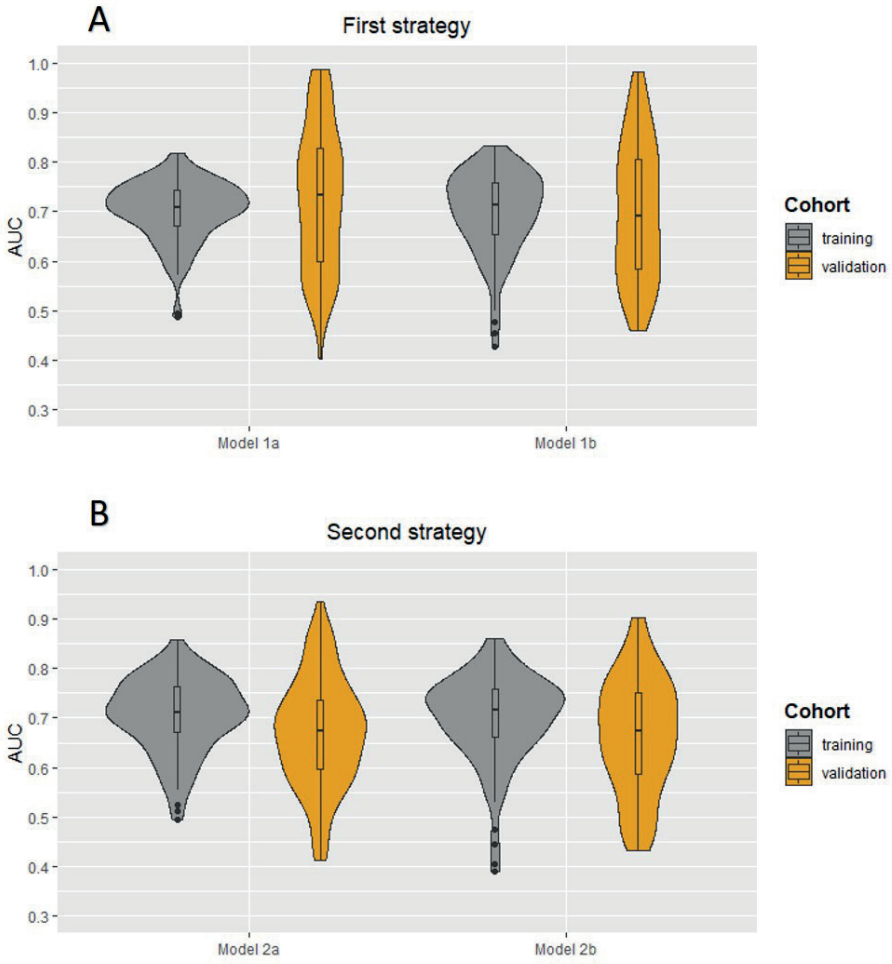


Figure S3. Violin plots for the radiomics models developed using the first (A) and second (B) strategy with additional feature se-lection step (ICC > 0.90): AUC value distributions (100 iterations) for the four models (1a, 1b, 2a and 2b) in both the training and validation cohort.

Table S1. The diagnostic performance of the radiomics models (100 iterations) with the exclusion of ROIs<50 voxels for the first and second strategy

Diagnostic parameters	First strategy								Second strategy							
	Model 1a								Model 2a							
	Training				Validation				Training				Validation			
	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)
Minimum	21	50	27	52	0	58	0	98	0	48	0	52	0	0	0	0
Median	62	70	64	67	50	75	2	99	62	69	64	67	39	67	56	64
Maximum	86	86	79	82	100	92	21	100	82	90	81	80	100	100	100	100

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; ROI, region of interest; sens, sensitivity; spec, specificity

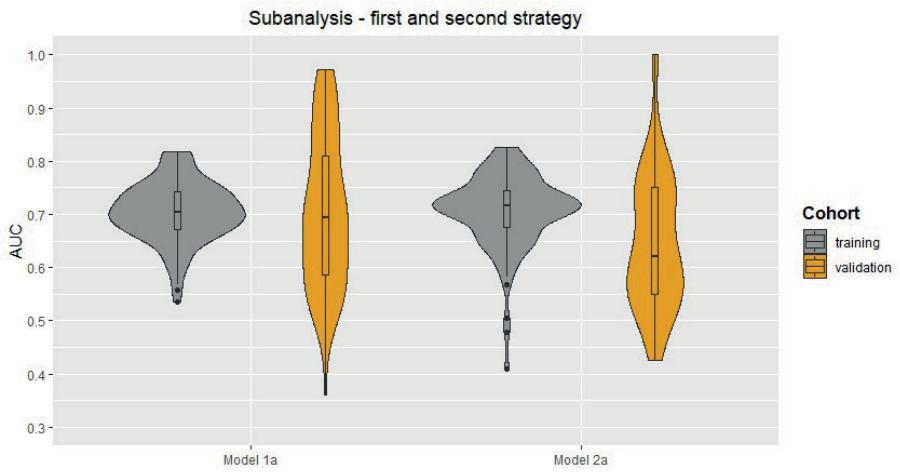


Figure S4. Violin plots for the radiomics models with the exclusion of ROIs<50 voxels developed using the first strategy and second strategy: AUC value distribution (100 iterations) for the two models (1a and 2a) in both the training and validation cohort. Abbreviations: ROI, region of interest

Clinical model development

The following clinical characteristics were available and selected for the development of the clinical models: patient age, clinical tumor size, clinical tumor stage, tumor histology, tumor grade, and receptor subtype (ER, PR, and HER2+). No highly correlated clinical characteristics were present. The minimum and maximum AUC values in the training cohorts were 0.52–0.66, 0.43–0.71, 0.41–0.67, and 0.43–0.74 for models 1a, 1b, 2a, and 2b, respectively. The median AUC values for all models in the training cohorts were between 0.59–0.60. All models showed a wider range of AUC values in the validation cohorts. The AUC value distribution for all models in the training and validation cohorts are presented in the violin plots in Figure 5. The minimum and maximum sensitivity in the training cohorts were 18–64%, 31–71%, 0–65%, and 33–73% for models 1a, 1b, 2a, and 2b, respectively. The median sensitivity for all models in the training cohorts was between 42–58%. All models showed lower median sensitivity in the validation cohorts, except for model 2b. The minimum and maximum positive predictive value (PPV) in the training cohorts were 42–71%, 41–85%, 48–73%, and 43–86% for models 1a, 1b, 2a, and 2b, respectively. The median PPV for all models in the training cohorts was between 68–70%. All models showed a lower median PPV in the validation cohorts, except for model 2a. In all four models, the clinical tumor size was ranked as the most important clinical characteristic followed by age. The diagnostic performance parameters of the clinical models (100 iterations) are shown in Table 3.

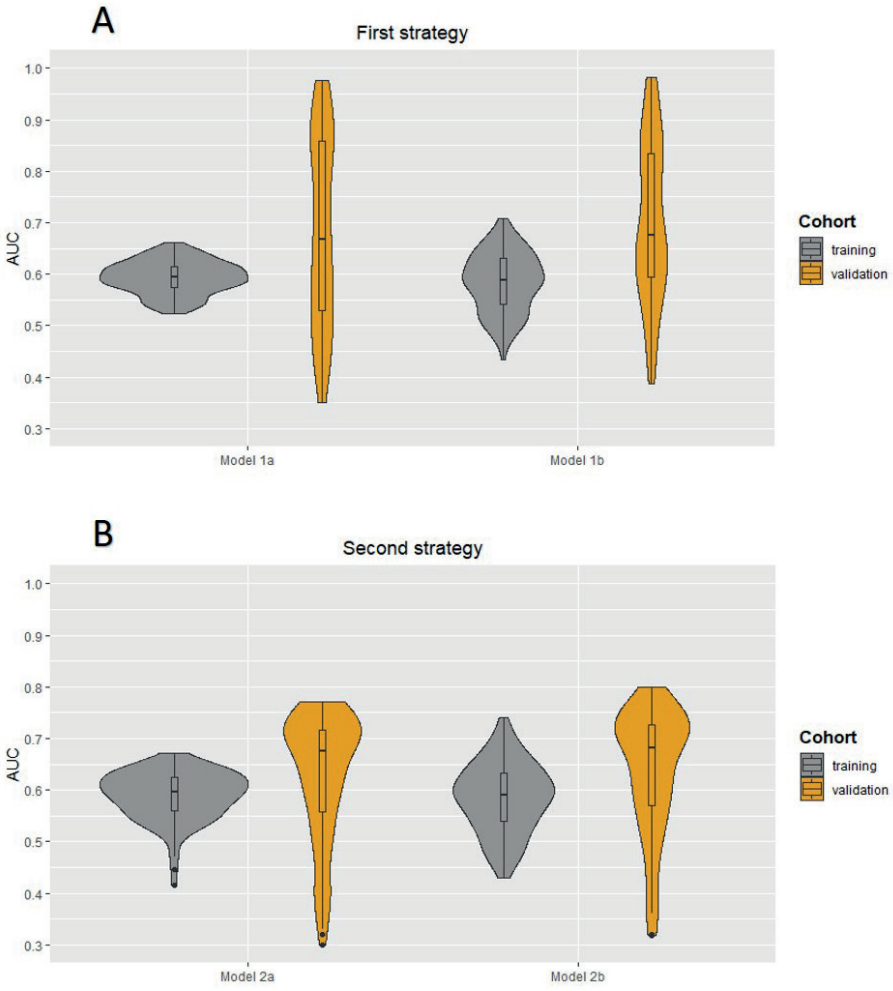


Figure 5. Violin plots for the clinical models developed using the first (***A***) and second (***B***) strategy: AUC value distributions (100 iterations) for the four models (1a, 1b, 2a, and 2b) in both the training and validation cohort.

Table 3. The diagnostic performance of the clinical models (100 iterations) for the first and second strategy.

Diagnostic parameters	Training				Validation				Training				Validation			
	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)
	First Strategy															
	Model 1a								Model 1b							
Minimum	18	64	42	65	0	40	0	99	31	46	41	42	0	14	0	97
Median	50	86	68	72	0	91	0	99	58	74	70	64	50	64	1	99
Maximum	64	93	71	78	100	99	18	100	71	92	85	73	100	88	9	100
	Second Strategy															
	Model 2a								Model 2b							
Minimum	0	55	48	61	0	0	10	34	33	45	43	43	0	0	10	0
Median	42	85	68	72	39	80	69	73	57	75	70	63	61	53	43	67
Maximum	65	100	73	80	100	100	73	84	73	91	86	74	100	100	100	86

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; sens, sensitivity; spec, specificity.

RQS and TRIPOD

This study scored a radiomics quality score (RQS) of 58% (21 out of 36 points) (Table S2). The score of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist was 67% (18 out of 27 applicable items).

Table S2. Radiomics Quality Score

Criteria	Points
Image protocol quality	+ 1
Multiple segmentations	+ 1
Phantom study on all scanners	+ 0
Imaging at multiple time points	+ 0
Feature reduction or adjustment for multiple testing	+ 3
Multivariate analysis with non radiomics features	+ 0
Detect and discuss biological correlates	+ 0
Cut-off analyses	+ 0
Discrimination statistics	+ 2
Calibration statistics	+ 1
Prospective study registered in a trial database	+ 7
Validation	+ 2
Comparison to 'gold standard'	+ 2
Potential clinical utility	+ 2
Cost-effectiveness analysis	+ 0
Open science and data	+ 0
Total	**21**

A total of 36 points can be achieved, with higher scores indicating higher research quality.

Discussion

Accurate preoperative prediction of axillary lymph node metastasis can assist in clinical decision-making regarding the extent of axillary surgery and radiation therapy, and provide essential prognostic information. In this study, clinical models and radiomics models based on T2-weighted dedicated axillary MRI features with node-by-node analysis were investigated for the preoperative prediction of axillary lymph node metastasis. The different sets of features selected at each split resulted in a wide range of AUC values and did not allow for the development of a final radiomics prediction model. The performance of the clinical models (AUC values between 0.41–0.74) was lower compared to the radiomics models (AUC values between 0.48–0.89) in the training cohorts. The validation results of both models showed a wider range of diagnostic performance parameters compared to the training results possibly explained by the small dataset, the methodology used for selection and model building, and potential overfitting. The wide AUC range in the clinical models leads us to the hypothesis that the small dataset contains unseen biological covariates, and that therefore the wide AUC range in the radiomics models cannot be explained by variations in imaging alone.

To the best of our knowledge, this is the first study investigating the role of MRI-based radiomics for the prediction of axillary lymph node metastasis in breast cancer patients by extracting features from delineated axillary lymph nodes. Previously published articles investigated the same topic by extracting the features from the delineated breast tumor[15,27,28]. These articles showed promising validation results with AUC values between 0.77–0.82. In this recent study, initially, the small ROI volumes were seen as a reason for the inconclusive results. If an ROI contains a low number of voxels, it may not be possible to calculate meaningful radiomics features[29]. However, after the subanalysis excluding ROI volumes less than 50 voxels, the AUC values were between 0.53–0.82 and 0.41–0.83 for the training cohorts for models 1a and 2a, respectively, which highlights the effects of differences in scan acquisition and reconstruction parameters. Furthermore, the skewed data in this recent study may have caused inconsistent results compared to the previous studies as models tend to favor the more common outcome.

To date, only two previously published articles extracted features from delineated lymph nodes for radiomics and deep learning analyses. The first article used a neural network to develop prediction models in head and neck cancer[30]. The second article developed a radiomics model based on CT images of colorectal cancer patients[31]. Both studies showed that there is potential by delineating lymph nodes for radiomics and deep learning analysis for the classification of positive and negative lymph nodes. The differences in results compared to this recent study may be due to the variety of implementation of the different steps in the radiomics workflow and the chosen imaging modality (CT vs. MRI).

The diagnostic performance of dedicated axillary T2W MRI for axillary lymph node staging has previously been investigated using node-by-node analysis[12]. Schipper et al. showed AUC values between 0.78–0.88, with a good interobserver agreement (kappa = 0.70). The current analysis with MRI-based radiomics using dedicated axillary T2W MR images suggested that the quantitative analysis did not exceed the qualitative analysis by the radiologists. It was decided to only perform radiomics analyses using the T2W MR images, as previous research indicated that diffusion-weighted images and apparent diffusion coefficient measurements have no added value for the axillary lymph node staging[12,32]. Furthermore, a recently published article has shown that the evaluation of axillary lymph nodes with dedicated axillary MRI is comparable to standard breast MRI with a complete field of view of the axillary region[32]. However, the majority of the breast MRI examinations are still performed with an incomplete field of view of the axillary region[9]. In addition, the coronal view of the dedicated axillary MRI possibly provides more accurate delineations compared to the transversal view of the standard breast MRI, which could be of added value to the radiomics analysis.

Most radiomics studies suffer from small and heterogeneous datasets collected from different imaging systems. In this current study, a great advantage for the radiomics analyses was the prospectively collected set of MR images on the same MRI scanner using an equal acquisition protocol with the patients in corresponding positions. Despite the prospectively collected dataset, a number of acquisition and reconstruction parameters varied depending on the patient. Furthermore, the different sets of features selected in every training cohort resulted in a wide range of AUC values and did not allow the development of a final radiomics prediction model. This could be justified by two theories: (i) The variations in acquisition and reconstruction parameters significantly affected the value of radiomics features, resulting in non-comparable data points; or (ii) Radiomics features do not have an added value in the prediction of axillary lymph nodes metastasis. However, theory (ii) is less likely, as radiomics models performed well in some splits. Future MRI phantom and reproducibility studies should investigate the effect of MR image acquisition and reconstruction parameters on feature values to determine repeatable and reproducible features. We nevertheless believe that it is also important to publish inconclusive radiomics results since publication bias seems to play a role in this research field, with only 6% of the radiomics articles presenting negative results[33].

This study also has certain limitations. The large skewness of the data with only 7% positive axillary lymph nodes was a drawback for the analyses. The skewness of the data was addressed by splitting the dataset using two different strategies and by using repeated cross-validation in the training cohort. However, it is important to note that the ratio of node-positive (19%) and node-negative (81%) breast

cancer patients in this study is comparable to the clinics. Besides the skewness of the data, the included number of patients was relatively low for radiomics analysis and selecting only node-positive patients in strategy 2 decreased the number even further. However, since the dedicated axillary MRI is not included in the breast MRI protocol and no similar public dataset is available, it is not possible to expand this current dataset. Lastly, manual delineation of the axillary lymph nodes was performed by one researcher, which potentially could be a major limitation of the findings because of the susceptibility of inter- and intra-observer variabilities[34]. Although this issue has been addressed in this current study by developing models based on only robust features for varying breast tumor delineations[23]. Based on the assumption that breast and lymph node delineations on MRI are comparable, varying delineations did not affect the radiomics results. However, this topic needs to be thoroughly investigated in future studies.

Conclusions

In conclusion, based on our results dedicated axillary MRI-based radiomics with node-by-node analysis did not contribute to the prediction of axillary lymph node metastasis based on data where variations in acquisition and reconstruction parameters were not addressed. Larger datasets combined with MRI phantom data and reproducibility studies are necessary to determine if further radiomics research using dedicated axillary MR images for the prediction of axillary lymph node metastasis is of added value.

Supplementary Material A

1 - No ICC Model 1a		2 - No ICC Model 1b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_Maximum2DDiameterColumn	98	firstorder_10Percentile	92
firstorder_10Percentile	91	shape_Maximum2DDiameterColumn	82
shape_LeastAxisLength	89	firstorder_Median	80
firstorder_90Percentile	89	firstorder_90Percentile	80
firstorder_Median	89	glszm_SizeZoneNonUniformityNormalized	74
glszm_LargeAreaLowGrayLevelEmphasis	83	firstorder_Kurtosis	71
gldm_LargeDependenceLowGrayLevelEmphasis	80	firstorder_Minimum	69
shape_MajorAxisLength	77	shape_LeastAxisLength	68
firstorder_Kurtosis	76	shape_MajorAxisLength	65
firstorder_Minimum	70	glcm_Idmn	60
glszm_SizeZoneNonUniformityNormalized	66	firstorder_RobustMeabsoluteDeviation	57
ngtdm_Busyness	63	glszm_LargeAreaLowGrayLevelEmphasis	55
shape_SurfaceVolumeRatio	60	shape_Sphericity	50
shape_Elongation	59	gldm_LargeDependenceLowGrayLevelEmphasis	48
firstorder_RobustMeabsoluteDeviation	54	firstorder_Maximum	46
ngtdm_Contrast	47	ngtdm_Contrast	45
shape_Sphericity	42	ngtdm_Busyness	45
firstorder_Maximum	35	shape_Elongation	42
glcm_Idmn	27	gldm_SmallDependenceEmphasis	37
shape_MinorAxisLength	26	firstorder_Skewness	26
shape_Flatness	25	shape_SurfaceVolumeRatio	21
glszm_LargeAreaHighGrayLevelEmphasis	21	gldm_DependenceVariance	20
firstorder_Skewness	16	firstorder_Energy	16
glcm_DifferenceVariance	10	shape_MinorAxisLength	16
gldm_DependenceVariance	7	glcm_Correlation	14
gldm_SmallDependenceEmphasis	6	shape_Flatness	14
firstorder_Energy	6	glcm_Imc2	12
gldm_SmallDependenceHighGrayLevelEmphasis	6	glcm_DifferenceVariance	12
glcm_Autocorrelation	5	ngtdm_Strength	11
ngtdm_Strength	4	glcm_Id	9
glcm_Correlation	4	gldm_SmallDependenceHighGrayLevelEmphasis	9

3 - No ICC Model 2a		4 - No ICC Model 2b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_Maximum2DDiameterColumn	79	firstorder_10Percentile	74
glszm_LargeAreaLowGrayLevelEmphasis	72	firstorder_Median	74
firstorder_10Percentile	69	firstorder_90Percentile	71
shape_MajorAxisLength	68	shape_Maximum2DDiameterColumn	70
shape_LeastAxisLength	67	firstorder_Kurtosis	63
firstorder_Median	67	glszm_SizeZoneNonUniformityNormalized	62
gldm_LargeDependenceLowGrayLevelEmphasis	65	shape_MajorAxisLength	59
firstorder_90Percentile	59	shape_LeastAxisLength	58
firstorder_Kurtosis	57	glszm_LargeAreaLowGrayLevelEmphasis	54
shape_Elongation	49	firstorder_Minimum	51
ngtdm_Busyness	48	glcm_Idmn	47
glszm_SizeZoneNonUniformityNormalized	44	ngtdm_Contrast	44
firstorder_RobustMeabsoluteDeviation	43	firstorder_RobustMeabsoluteDeviation	43
firstorder_Minimum	43	gldm_LargeDependenceLowGrayLevelEmphasis	42
ngtdm_Contrast	33	ngtdm_Busyness	39
firstorder_Maximum	32	shape_Sphericity	39
shape_SurfaceVolumeRatio	30	shape_Elongation	38
shape_Sphericity	26	firstorder_Maximum	36
shape_Flatness	26	firstorder_Skewness	24
glcm_Idmn	24	glcm_Imc2	21
shape_MinorAxisLength	19	gldm_SmallDependenceEmphasis	21
firstorder_Skewness	19	shape_SurfaceVolumeRatio	20
glcm_DifferenceVariance	16	gldm_SmallDependenceHighGrayLevelEmphasis	20
gldm_SmallDependenceHighGrayLevelEmphasis	14	glcm_DifferenceVariance	19
glszm_LargeAreaHighGrayLevelEmphasis	12	shape_MinorAxisLength	14
ngtdm_Strength	11	firstorder_Energy	11
firstorder_Energy	9	glcm_Id	11
glcm_Imc2	8	shape_Flatness	10
glcm_Autocorrelation	7	gldm_DependenceVariance	9
gldm_DependenceVariance	6	ngtdm_Strength	9
gldm_SmallDependenceEmphasis	5	glcm_ClusterShade	7

1 - No ICC Model 1a		2 - No ICC Model 1b	
Feature	**Iteration**	**Feature**	**Iteration**
glcm_Imc2	3	glszm_LargeAreaHighGrayLevelEmphasis	8
gldm_GrayLevelNonUniformity	3	firstorder_Range	8
glszm_LowGrayLevelZoneEmphasis	2	glszm_LowGrayLevelZoneEmphasis	6
ngtdm_Coarseness	1	glcm_Autocorrelation	6
glcm_Imc1	1	glcm_InverseVariance	6
shape_Maximum2DDiameterRow	1	gldm_DependenceNonUniformityNormalized	6
		glcm_Imc1	5
		glrlm_LongRunLowGrayLevelEmphasis	5
		gldm_GrayLevelNonUniformity	3
		gldm_LargeDependenceHighGrayLevelEmphasis	3
		glcm_ClusterShade	3
		glszm_GrayLevelVariance	1
		glcm_ClusterProminence	1
		glszm_SmallAreaLowGrayLevelEmphasis	1

5 - ICC 0.75 Model 1a		6 - ICC 0.75 Model 1b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_Maximum2DDiameterColumn	97	firstorder_Median	88
shape_LeastAxisLength	93	shape_Maximum2DDiameterColumn	86
firstorder_90Percentile	93	firstorder_90Percentile	86
firstorder_Median	90	glszm_SizeZoneNonUniformityNormalized	84
gldm_LargeDependenceLowGrayLevelEmphasis	90	shape_LeastAxisLength	81
glszm_LargeAreaLowGrayLevelEmphasis	89	firstorder_RobustMeabsoluteDeviation	75
shape_MajorAxisLength	89	glcm_Idmn	70
glszm_SizeZoneNonUniformityNormalized	85	shape_MajorAxisLength	68
glszm_SizeZoneNonUniformity	76	firstorder_Maximum	67
ngtdm_Busyness	73	shape_Sphericity	67
shape_SurfaceVolumeRatio	73	gldm_LargeDependenceLowGrayLevelEmphasis	66
firstorder_RobustMeabsoluteDeviation	72	glszm_LargeAreaLowGrayLevelEmphasis	65

3 - No ICC Model 2a		4 - No ICC Model 2b	
Feature	**Iteration**	**Feature**	**Iteration**
glcm_ClusterShade	5	glcm_Imc1	7
glcm_Correlation	5	firstorder_Range	6
glszm_LowGrayLevelZoneEmphasis	3	glcm_Correlation	5
shape_Maximum2DDiameterRow	3	glcm_Autocorrelation	5
glcm_Imc1	3	glcm_InverseVariance	4
glrlm_LongRunLowGrayLevelEmphasis	2	gldm_GrayLevelNonUniformity	4
firstorder_Range	2	ngtdm_Complexity	3
glcm_SumEntropy	1	firstorder_Uniformity	3
firstorder_Uniformity	1	gldm_ DependenceNonUniformityNormalized	3
glcm_JointEnergy	1	shape_Maximum2DDiameterRow	3
glszm_GrayLevelVariance	1	glrlm_LongRunLowGrayLevelEmphasis	2
gldm_GrayLevelNonUniformity	1	glszm_SmallAreaLowGrayLevelEmphasis	2
		ngtdm_Coarseness	2
		glszm_LowGrayLevelZoneEmphasis	1
		glszm_LargeAreaHighGrayLevelEmphasis	1
		glcm_SumEntropy	1
		glcm_DifferenceEntropy	1
		glcm_ClusterProminence	1
		gldm_LargeDependenceHighGrayLevelEmphasis	1

7 - ICC 0.75 Model 2a		8 - ICC 0.75 Model 2b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_Maximum2DDiameterColumn	88	firstorder_Median	81
glszm_LargeAreaLowGrayLevelEmphasis	81	firstorder_90Percentile	79
shape_LeastAxisLength	80	shape_Maximum2DDiameterColumn	76
firstorder_Median	78	shape_MajorAxisLength	71
gldm_LargeDependenceLowGrayLevelEmphasis	76	glszm_SizeZoneNonUniformityNormalized	69
shape_MajorAxisLength	75	shape_LeastAxisLength	64
firstorder_90Percentile	74	glszm_LargeAreaLowGrayLevelEmphasis	63
firstorder_RobustMeabsoluteDeviation	61	firstorder_RobustMeabsoluteDeviation	58
glszm_SizeZoneNonUniformityNormalized	60	gldm_LargeDependenceLowGrayLevelEmphasis	58
firstorder_Maximum	57	firstorder_Maximum	52
ngtdm_Busyness	55	glcm_Idmn	48
shape_Sphericity	52	ngtdm_Contrast	47

5 - ICC 0.75 Model 1a		6 - ICC 0.75 Model 1b	
Feature	**Iteration**	**Feature**	**Iteration**
firstorder_Maximum	72	ngtdm_Contrast	55
shape_Sphericity	61	ngtdm_Busyness	48
ngtdm_Contrast	58	glszm_SizeZoneNonUniformity	46
glcm_Idmn	50	gldm_SmallDependenceEmphasis	42
shape_Flatness	47	firstorder_Skewness	34
firstorder_Skewness	40	shape_Flatness	29
shape_MinorAxisLength	34	shape_SurfaceVolumeRatio	25
glszm_LargeAreaHighGrayLevelEmphasis	26	firstorder_Range	21
gldm_SmallDependenceEmphasis	15	gldm_DependenceVariance	21
ngtdm_Strength	15	shape_MinorAxisLength	20
glcm_DifferenceVariance	13	firstorder_Energy	19
gldm_DependenceVariance	11	glcm_DifferenceVariance	15
gldm_SmallDependenceHighGrayLevelEmphasis	11	glcm_Id	14
firstorder_Energy	8	glcm_Imc2	14
glcm_Autocorrelation	8	gldm_SmallDependenceHighGrayLevelEmphasis	13
glrlm_RunLengthNonUniformity	4	glcm_Autocorrelation	13
glcm_Imc2	3	glszm_LargeAreaHighGrayLevelEmphasis	12
glcm_Id	3	ngtdm_Strength	12
firstorder_Range	3	glrlm_RunLengthNonUniformity	10
shape_Maximum2DDiameterRow	3	glcm_InverseVariance	9
glcm_Imc1	2	gldm_ DependenceNonUniformityNormalized	6
gldm_ DependenceNonUniformityNormalized	1	glcm_ClusterTendency	6
gldm_GrayLevelNonUniformity	1	gldm_GrayLevelNonUniformity	3
glcm_ClusterTendency	1	glcm_DifferenceEntropy	3
		glcm_MaximumProbability	3
		glcm_Imc1	2
		glszm_GrayLevelVariance	2
		glcm_SumEntropy	1
		ngtdm_Complexity	1

7 - ICC 0.75 Model 2a		8 - ICC 0.75 Model 2b	
Feature	**Iteration**	**Feature**	**Iteration**
glszm_SizeZoneNonUniformity	50	shape_Sphericity	47
shape_Flatness	47	ngtdm_Busyness	42
shape_SurfaceVolumeRatio	45	firstorder_Skewness	41
ngtdm_Contrast	44	glszm_SizeZoneNonUniformity	39
firstorder_Skewness	39	gldm_SmallDependenceEmphasis	26
glcm_Idmn	36	glcm_DifferenceVariance	26
shape_MinorAxisLength	28	shape_Flatness	25
glcm_DifferenceVariance	24	shape_SurfaceVolumeRatio	23
firstorder_Energy	19	glcm_Imc2	22
ngtdm_Strength	15	gldm_SmallDependenceHighGrayLevelEmphasis	21
gldm_DependenceVariance	15	shape_MinorAxisLength	18
gldm_SmallDependenceHighGrayLevelEmphasis	15	firstorder_Range	18
glcm_Imc2	12	gldm_DependenceVariance	17
glszm_LargeAreaHighGrayLevelEmphasis	11	firstorder_Energy	15
glcm_Autocorrelation	11	glcm_Id	13
firstorder_Range	11	ngtdm_Strength	11
shape_Maximum2DDiameterRow	10	glcm_InverseVariance	11
glcm_ClusterTendency	10	glszm_LargeAreaHighGrayLevelEmphasis	8
gldm_SmallDependenceEmphasis	8	glcm_Imc1	8
glcm_Imc1	5	glcm_Autocorrelation	8
glszm_GrayLevelVariance	4	gldm_ DependenceNonUniformityNormalized	7
glrlm_RunLengthNonUniformity	3	glcm_DifferenceEntropy	5
glcm_SumEntropy	3	glrlm_RunLengthNonUniformity	4
ngtdm_Complexity	2	ngtdm_Complexity	4
gldm_GrayLevelNonUniformity	2	glszm_GrayLevelVariance	4
glrlm_GrayLevelNonUniformityNormalized	2	glcm_ClusterTendency	3
gldm_ DependenceNonUniformityNormalized	1	gldm_GrayLevelNonUniformity	3
glcm_InverseVariance	1	shape_Maximum2DDiameterRow	3
glcm_Id	1	ngtdm_Coarseness	2
glcm_MaximumProbability	1	glcm_SumEntropy	1

9 - ICC 0.8 Model 1a		10 - ICC 0.8 Model 1b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_Maximum2DDiameterColumn	99	shape_Maximum2DDiameterColumn	89
firstorder_90Percentile	93	firstorder_90Percentile	87
glszm_LargeAreaLowGrayLevelEmphasis	92	glszm_SizeZoneNonUniformityNormalized	84
shape_LeastAxisLength	92	firstorder_Median	84
firstorder_Median	91	shape_LeastAxisLength	81
gldm_LargeDependenceLowGrayLevelEmphasis	90	shape_MajorAxisLength	78
shape_MajorAxisLength	89	gldm_LargeDependenceLowGrayLevelEmphasis	69
glszm_SizeZoneNonUniformityNormalized	87	glszm_LargeAreaLowGrayLevelEmphasis	69
glszm_SizeZoneNonUniformity	85	firstorder_Maximum	68
shape_SurfaceVolumeRatio	77	ngtdm_Busyness	52
firstorder_Maximum	75	glszm_SizeZoneNonUniformity	51
ngtdm_Busyness	73	shape_SurfaceVolumeRatio	47
shape_Flatness	68	gldm_SmallDependenceEmphasis	43
shape_MinorAxisLength	58	shape_Flatness	42
glszm_LargeAreaHighGrayLevelEmphasis	48	glrlm_RunLengthNonUniformityNormalized	41
glrlm_RunLengthNonUniformityNormalized	44	glcm_DifferenceVariance	37
glcm_DifferenceVariance	34	glcm_Imc2	30
glrlm_ShortRunHighGrayLevelEmphasis	33	glcm_Id	30
gldm_SmallDependenceEmphasis	25	glrlm_ShortRunHighGrayLevelEmphasis	28
gldm_DependenceVariance	20	gldm_DependenceVariance	25
firstorder_TotalEnergy	14	glcm_InverseVariance	24
glcm_Imc2	14	shape_MinorAxisLength	23
firstorder_Range	13	firstorder_Range	22
shape_Maximum2DDiameterRow	13	firstorder_TotalEnergy	21
glcm_SumEntropy	9	glszm_LargeAreaHighGrayLevelEmphasis	16
glcm_Imc1	8	gldm_DependenceNonUniformityNormalized	11
glcm_InverseVariance	6	glcm_Imc1	10
glcm_Id	6	glcm_SumEntropy	10
glcm_DifferenceEntropy	4	glszm_GrayLevelVariance	6
glszm_GrayLevelVariance	2	gldm_GrayLevelNonUniformity	4
gldm_DependenceNonUniformityNormalized	1	shape_Maximum2DDiameterRow	4
gldm_GrayLevelNonUniformity	1	glszm_SmallAreaHighGrayLevelEmphasis	3
ngtdm_Complexity	1	ngtdm_Complexity	2
glszm_SmallAreaHighGrayLevelEmphasis	1		

11 - ICC 0.8 Model 2a		12 - ICC 0.8 Model 2b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_Maximum2DDiameterColumn	90	shape_Maximum2DDiameterColumn	84
shape_LeastAxisLength	87	firstorder_Median	82
shape_MajorAxisLength	86	firstorder_90Percentile	79
gldm_LargeDependenceLowGrayLevelEmphasis	84	shape_LeastAxisLength	79
glszm_LargeAreaLowGrayLevelEmphasis	83	shape_MajorAxisLength	74
firstorder_Median	81	glszm_SizeZoneNonUniformityNormalized	72
firstorder_90Percentile	74	glszm_LargeAreaLowGrayLevelEmphasis	69
shape_Flatness	69	gldm_LargeDependenceLowGrayLevelEmphasis	58
glszm_SizeZoneNonUniformityNormalized	68	firstorder_Maximum	53
ngtdm_Busyness	64	ngtdm_Busyness	51
firstorder_Maximum	63	glszm_SizeZoneNonUniformity	51
shape_SurfaceVolumeRatio	61	shape_Flatness	36
glszm_SizeZoneNonUniformity	59	glcm_DifferenceVariance	35
shape_MinorAxisLength	49	shape_SurfaceVolumeRatio	34
glcm_DifferenceVariance	45	glrlm_RunLengthNonUniformityNormalized	32
glrlm_RunLengthNonUniformityNormalized	35	glcm_Imc2	31
glrlm_ShortRunHighGrayLevelEmphasis	30	firstorder_Range	31
gldm_DependenceVariance	28	gldm_SmallDependenceEmphasis	28
glszm_LargeAreaHighGrayLevelEmphasis	26	glrlm_ShortRunHighGrayLevelEmphasis	28
glcm_Imc2	23	glcm_Id	26
firstorder_Range	22	shape_MinorAxisLength	25
shape_Maximum2DDiameterRow	22	gldm_DependenceVariance	23
gldm_SmallDependenceEmphasis	18	firstorder_TotalEnergy	19
firstorder_TotalEnergy	18	glcm_Imc1	18
glszm_GrayLevelVariance	15	glcm_InverseVariance	17
glcm_Imc1	14	glcm_SumEntropy	12
glcm_SumEntropy	14	glszm_GrayLevelVariance	12
glcm_Id	14	shape_Maximum2DDiameterRow	11
glszm_SmallAreaHighGrayLevelEmphasis	8	glcm_DifferenceEntropy	10
glcm_DifferenceEntropy	6	glszm_LargeAreaHighGrayLevelEmphasis	9
gldm_GrayLevelNonUniformity	5	gldm_DependenceNonUniformityNormalized	8
glcm_InverseVariance	5	glszm_SmallAreaHighGrayLevelEmphasis	8
ngtdm_Complexity	5	ngtdm_Complexity	5
glrlm_GrayLevelNonUniformityNormalized	2	glrlm_GrayLevelNonUniformityNormalized	4
gldm_DependenceNonUniformityNormalized	1	gldm_GrayLevelNonUniformity	3

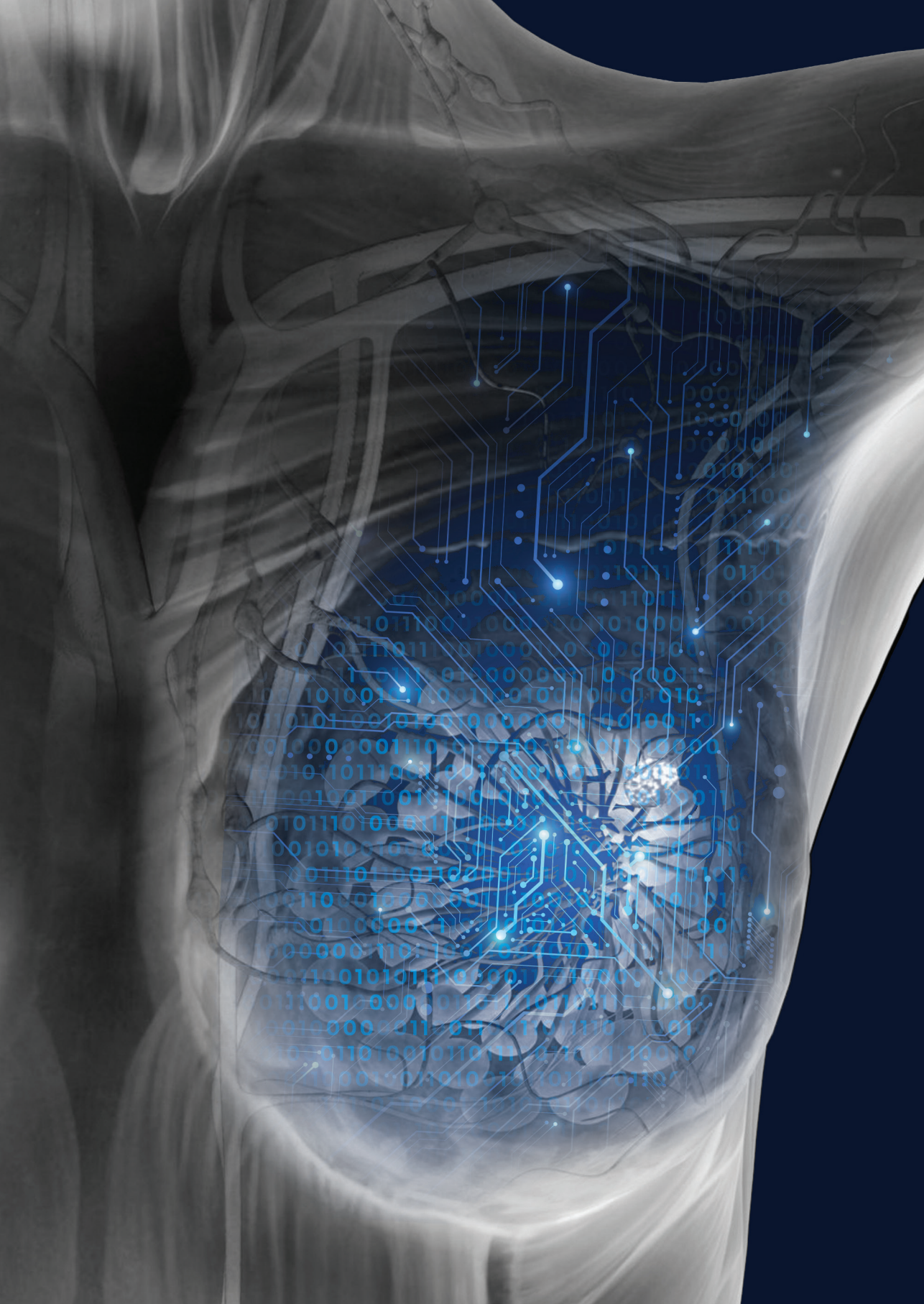
13 - ICC 0.9 Model 1a		14 - ICC 0.9 Model 1b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_LeastAxisLength	100	shape_LeastAxisLength	97
glszm_LargeAreaHighGrayLevelEmphasis	93	firstorder_90Percentile	88
firstorder_90Percentile	92	glszm_GrayLevelNonUniformity	87
glszm_GrayLevelNonUniformity	87	glszm_LargeAreaHighGrayLevelEmphasis	85
glszm_LargeAreaLowGrayLevelEmphasis	83	glszm_LargeAreaLowGrayLevelEmphasis	72
ngtdm_Busyness	62	ngtdm_Busyness	64
firstorder_Energy	39	firstorder_Energy	55
firstorder_Maximum	35	firstorder_Maximum	44

15 - ICC 0.9 Model 2a		16 - ICC 0.9 Model 2b	
Feature	**Iteration**	**Feature**	**Iteration**
shape_LeastAxisLength	98	shape_LeastAxisLength	93
glszm_LargeAreaHighGrayLevelEmphasis	84	firstorder_90Percentile	85
firstorder_90Percentile	84	glszm_GrayLevelNonUniformity	84
glszm_GrayLevelNonUniformity	83	glszm_LargeAreaHighGrayLevelEmphasis	76
glszm_LargeAreaLowGrayLevelEmphasis	76	glszm_LargeAreaLowGrayLevelEmphasis	72
ngtdm_Busyness	63	ngtdm_Busyness	62
firstorder_Energy	50	firstorder_Energy	58
firstorder_Maximum	50	firstorder_Maximum	53

References

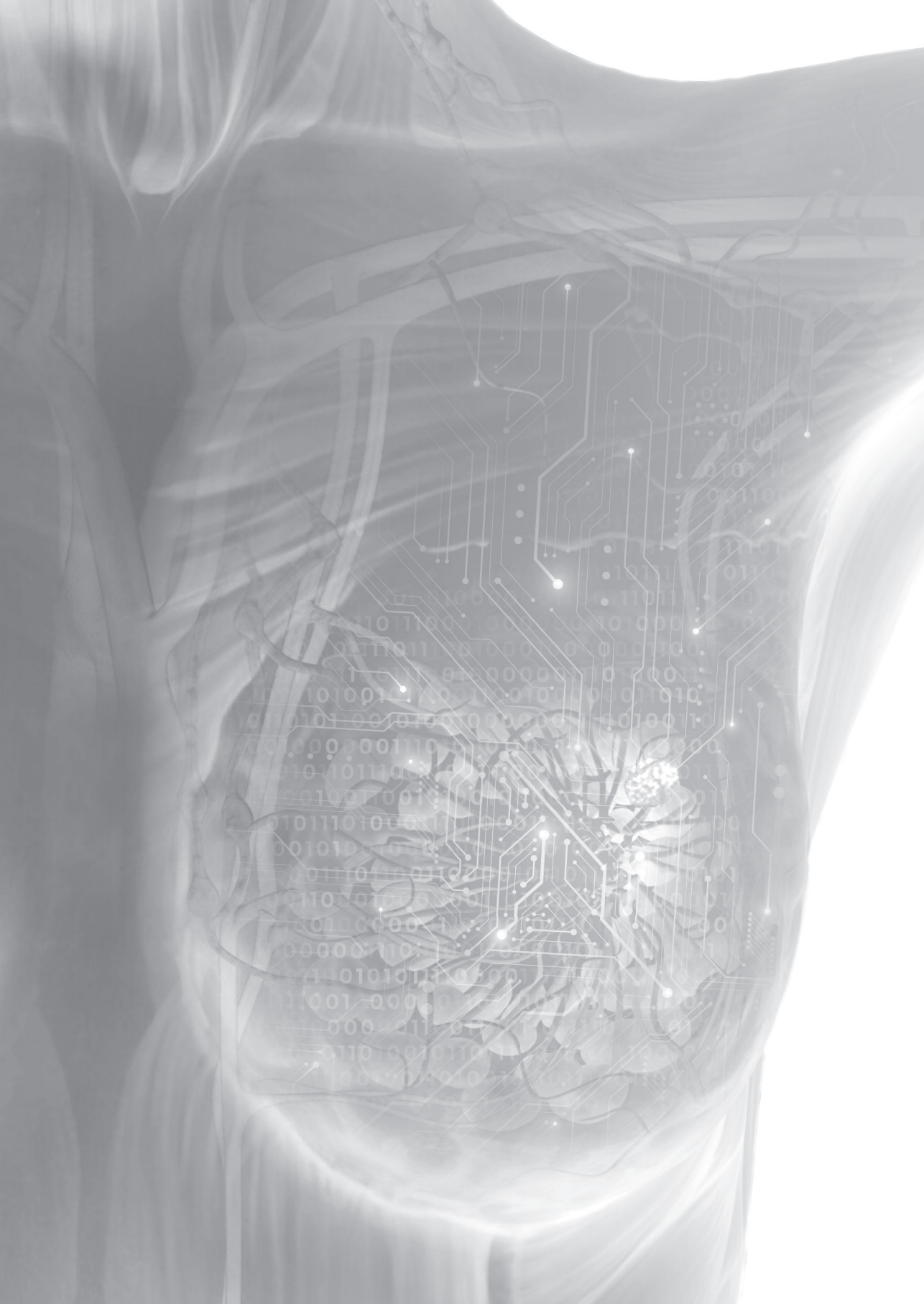
1. Beenken SW, Urist MM, Zhang Y, et al. Axillary lymph node status, but not tumor size, predicts locoregional recurrence and overall survival after mastectomy for breast cancer. *Annals of surgery.* 2003;237(5):732-738; discussion 738-739.
2. Soerjomataram I, Louwman MW, Ribot JG, Roukema JA, Coebergh JW. An overview of prognostic factors for long-term survivors of breast cancer. *Breast cancer research and treatment.* 2008;107(3):309-330.
3. Carter SA, Lyons GR, Kuerer HM, et al. Operative and Oncologic Outcomes in 9861 Patients with Operable Breast Cancer: Single-Institution Analysis of Breast Conservation with Oncoplastic Reconstruction. *Annals of surgical oncology.* 2016;23(10):3190-3198.
4. Fisher B, Bauer M, Wickerham DL, et al. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. *Cancer.* 1983;52(9):1551-1557.
5. Surveillance, Epidemiology, and End Results Program (SEER). Table 4.13: Cancer of the Female Breast (Invasive). 5-Year Relative and Period Survival by Race, Diagnosis Year, Age and Stage at Diagnosis. *In: SEER Cancer Statistics Review (CSR) 1975-2012*
6. Senkus E, Kyriakides S, Ohno S, et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* 2015;26 Suppl 5:v8-30.
7. Caudle AS, Cupp JA, Kuerer HM. Management of axillary disease. *Surg Oncol Clin N Am.* 2014;23(3):473-486.
8. Sardanelli F, Boetes C, Borisch B, et al. Magnetic resonance imaging of the breast: recommendations from the EUSOMA working group. *European journal of cancer (Oxford, England : 1990).* 2010;46(8):1296-1316.
9. van Nijnatten TJA, Ploumen EH, Schipper RJ, et al. Routine use of standard breast MRI compared to axillary ultrasound for differentiating between no, limited and advanced axillary nodal disease in newly diagnosed breast cancer patients. *European journal of radiology.* 2016;85(12):2288-2294.
10. Kvistad KA, Rydland J, Smethurst HB, Lundgren S, Fjosne HE, Haraldseth O. Axillary lymph node metastases in breast cancer: preoperative detection with dynamic contrast-enhanced MRI. *Eur Radiol.* 2000;10(9):1464-1471.
11. Murray AD, Staff RT, Redpath TW, et al. Dynamic contrast enhanced MRI of the axilla in women with breast cancer: comparison with pathology of excised nodes. *Br J Radiol.* 2002;75(891):220-228.
12. Schipper RJ, Paiman ML, Beets-Tan RGH, et al. Diagnostic Performance of Dedicated Axillary T2-and Diffusion-weighted MR Imaging for Nodal Staging in Breast Cancer. *Radiology.* 2015;275(2):345-355.
13. Gillies R, Kinahan P, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology.* 2016;278(2):563-577.
14. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990).* 2012;48(4):441-446.
15. Dong Y, Feng Q, Yang W, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. *Eur Radiol.* 2018;28(2):582-591.
16. Han L, Zhu Y, Liu Z, et al. Radiomic nomogram for prediction of axillary lymph node metastasis in breast cancer. *Eur Radiol.* 2019;29(7):3820-3829.

17. Yang J, Wang T, Yang L, et al. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer Using Mammography-Based Radiomics Method. *Sci Rep.* 2019;9(1):4429.
18. Yu FH, Wang JX, Ye XH, Deng J, Hang J, Yang B. Ultrasound-based radiomics nomogram: A potential biomarker to predict axillary lymph node metastasis in early-stage invasive breast cancer. *European journal of radiology.* 2019;119:108658.
19. Liu C, Ding J, Spuhler K, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced MRI. *J Magn Reson Imaging.* 2018.
20. van Nijnatten TJA, Schipper RJ, Lobbes MBI, et al. Diagnostic performance of gadofosveset-enhanced axillary MRI for nodal (re)staging in breast cancer patients: results of a validation study. *Clin Radiol.* 2018;73(2):168-175.
21. Schipper RJ, Smidt ML, van Roozendaal LM, et al. Noninvasive nodal staging in patients with breast cancer using gadofosveset-enhanced magnetic resonance imaging: a feasibility study. *Invest Radiol.* 2013;48(3):134-139.
22. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017;77(21):e104-e107.
23. Granzier RWY, Verbakel NMH, Ibrahim A, et al. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. *Sci Rep.* 2020;10(1):14163.
24. Racine JS. RStudio: a platform-independent IDE for R and Sweave. *Journal of Applied Econometrics.* 2012;27(1):167-172.
25. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749-762.
26. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
27. Chai R, Ma H, Xu M, et al. Differentiating axillary lymph node metastasis in invasive breast cancer patients: A comparison of radiomic signatures from multiparametric breast MR sequences. *J Magn Reson Imaging.* 2019;50(4):1125-1132.
28. Tan H, Gan F, Wu Y, et al. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Carcinoma Using Radiomics Features Based on the Fat-Suppressed T2 Sequence. *Acad Radiol.* 2019.
29. Court LE, Fave X, Mackin D, Lee J, Yang JZ, Zhang LF. Computational resources for radiomics. *Transl Cancer Res.* 2016;5(4):340-348.
30. Ho TY, Chao CH, Chin SC, Ng SH, Kang CJ, Tsang NM. Classifying Neck Lymph Nodes of Head and Neck Squamous Cell Carcinoma in MRI Images with Radiomic Features. *J Digit Imaging.* 2020;33(3):613-618.
31. Li M, Zhang J, Dan Y, et al. A clinical-radiomics nomogram for the preoperative prediction of lymph node metastasis in colorectal cancer. *J Transl Med.* 2020;18(1):46.
32. Samiei S, Smidt ML, Vanwetswinkel S, et al. Diagnostic performance of standard breast MRI compared to dedicated axillary MRI for assessment of node-negative and node-positive breast cancer. *European Radiology.* 2020;30(8):4212-4222.
33. Buvat I, Orlhac F. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. *J Nucl Med.* 2019;60(11):1543-1544.
34. Saha A, Grimm LJ, Harowicz M, et al. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. *Med Phys.* 2016;43(8):4558.



PART II

Optimization in MRI-based radiomics



CHAPTER 6

Test-retest data for the assessment of breast MRI radiomic feature repeatability

Renée W.Y. Granzier, Abdalla Ibrahim*, Sergey P. Primakov*, Simon A. Keek, Iva Halilaj, Alex Zwanenburg, Sanne M. E. Engelen, Marc B. I. Lobbes, Philip Lambin, Henry C. Woodruff H.C, Marjolein L. Smidt

*Shared authorship

J Magn Reson Imaging. 2021 Dec 22

Abstract

Background: Radiomic features extracted from breast MRI have potential for diagnostic, prognostic, and predictive purposes. However, before they can be used as biomarkers in clinical decision support systems, features need to be repeatable and reproducible.
Objective: Identify repeatable radiomics features within breast tissue on prospectively collected MRI exams through multiple test-retest measurements.
Study type: Prospective
Population: 11 healthy female volunteers
Field strength/sequence: 1.5 T; MRI exams, comprising T2-weighted turbo spin-echo (T2W) sequence, native T1-weighted turbo gradient-echo (T1W) sequence, diffusion-weighted imaging (DWI) sequence using b-values 0/150/800, and corresponding derived ADC maps.
Assessment: 18 MRI exams (three test-retest settings, repeated on two days) per healthy volunteer were examined on an identical scanner using a fixed clinical breast protocol. For each scan, 91 features were extracted from the 3D manually segmented right breast using Pyradiomics, before and after image pre-processing. Image pre-processing consisted of (i) bias field correction (BFC), (ii) z-score normalization with and without BFC, (iii) grayscale discretization using 32 and 64 bins with and without BFC, and (iv) z-score normalization + grayscale discretization using 32 and 64 bins with and without BFC.
Statistical tests: Features' repeatability was assessed using concordance correlation coefficient(CCC) for each pair, i.e. each MRI was compared to each of the remaining 17 MRI with a cut-off value of CCC>0.90.
Results: Images without pre-processing produced the highest number of repeatable features for both T1W sequence and ADC maps with 15/91 (16.5%) and 8/91 (8.8%) repeatable features, respectively. Pre-processed images produced between 4/91 (4.4%) and 14/91 (15.4%), and 6/91 (6.6%) and 7/91 (7.7%) repeatable features, respectively for T1W and ADC maps. Z-score normalization produced highest number of repeatable features, 26/91 (28.6%) in T2W sequences, in these images, no pre-processing produced 11/91 (12.1%) repeatable features.
Data conclusion: Radiomic features extracted from T1W, T2W sequences and ADC maps from breast MRI exams showed a varying number of repeatable features, depending on the sequence. Effects of different preprocessing procedures on repeatability of features were different for each sequence.

Introduction

The use of radiomics to answer diagnostic, predictive, and prognostic questions has increased in recent years, especially in the field of oncology [1]. Radiomics refers to the extraction of large amounts of high-throughput quantitative data from medical images using mathematical algorithms that have the potential to noninvasively reveal more information about the region of interest than can be captured by visual inspection alone [2]. The extracted quantitative data, termed radiomics features, capture information regarding the shape, intensity, and texture of the chosen region of interest (ROI), which is usually the lesion or the affected organ. Radiomics features are intended to serve as biomarkers for the development of clinical decision support systems to enhance personalized medicine [3].

In breast cancer research, multiple radiomics studies have shown promising results for diagnostic, prognostic, and predictive purposes [4-6]. Despite these seemingly promising results, translation to clinical practice is limited [7]. A major translational bottleneck can be attributed to the often unknown effect that multiple steps in the radiomics workflow have on feature values, including image acquisition, reconstruction, and pre-processing [8-11]. For a radiomics feature to serve as a biomarker, and to be used reliably in clinical decision support systems, it must fulfill the criteria *repeatability* and *reproducibility* [12]. Repeatability can be defined as "the variability of the biomarker when repeated measurements are acquired on the same experimental unit under identical or nearly identical conditions" and reproducibility as to "variability in the biomarker measurements associated with using the imaging instrument in real-world clinical settings, which are subject to a variety of external factors that cannot all be tightly controlled" [12].

Previous research has already identified several steps in the radiomics workflow that influence the reproducibility and repeatability of radiomics features. For example, image acquisition and reconstruction appear to cause variation in radiomic feature values in research performed on CT imaging [13,14]. Unlike the Hounsfield Units in CT, MRI does not have absolute signal intensities, potentially causing large differences between images, emphasizing the importance of inspecting and possibly adjusting image intensities before performing feature extraction [15]. A test-retest MRI study of glioblastoma showed that both normalization and intensity quantization strategies affect radiomic feature repeatability and that the optimal strategy must be composed per feature group [16]. Further test-retest studies assessing feature repeatability have been performed in cervical [17], and prostate cancer [18,19] and have shown consistent results, although all studies state that translation of results to other tumor sites has not been confirmed. In contrast, Peerlings et al. [20] showed that 9.2% (122/1322) of the features, extracted from apparent diffusion coefficient (ADC) maps in ovarian, liver, and colorectal cancer patients, were repeatable among the different tumor sites.

The assessment of radiomics feature repeatability by test-retest studies in breast MRI exams is currently lacking. A potential reason for this lack of data is the variance present in a standard clinical breast MRI protocol, which means that scanning parameters may differ between patients scanned with the same clinical protocol. Therefore, this study investigated the repeatability of radiomics feature values extracted from breast MRI exams using a fixed clinical breast protocol comprising of T2-weighted (T2W) images, T1-weighted (T1W) images, and diffusion-weighted images (DWI) and their derived ADC maps.

Material and methods

Study population

The study was approved by the local medical ethical committee and written informed consent was given by all participants before participation. Eleven healthy female volunteers were recruited via college-wide advertisement. Participants were only included if they did not suffer from claustrophobia and met the requirements for admission to the MRI. Participants' height, weight, and the phase of the menstrual cycle were noted. The menstrual cycle of the included healthy volunteers was not taken into account during the MRI exams

Imaging acquisition

All MRI exams were performed using a 16-channel breast coil on one single 1.5 Tesla scanner (Ingenia, Philips Healthcare, Best, The Netherlands) in the same research institution by the same technician. During imaging, the women lay in the prone position with both breasts in the openings of the breast coil and both arms above their head. The performed MRI protocol consisted of a T2-weighted turbo spin echo (T2W), native T1-weighted turbo gradient echo (T1W), and a single shot diffusion-weighted imaging (DWI) sequence using b-values of 0, 150, and 800. A single corresponding ADC-map was derived from all three DWI sequences. All volunteers underwent MRI exams using the identical breast protocol while maintaining as many parameters fixed as possible. The acquisition parameters for the different MRI sequences are shown in the supplementary material (Table S1). The shimbox, needed for the T1W and DWI sequences, was placed on the sternum by default. In case the technician judged the scan as clinically insufficient, the shimbox was placed on the breasts.

Study design

A test-retest study was designed to assess the repeatability of breast-MRI extracted radiomic features. Three separate test-retest strategies were performed twice at six to ten day intervals. From here on, we will use 'date 1' to refer to the first scanning date of each healthy volunteer and 'date 2' to refer to the second scanning date. In each strategy, the complete breast MRI protocol was repeated three times with a two-minute pause between each protocol. In the first strategy (S1) the participant remained in the MRI scanner the entire time (including the pauses) without movement, for the acquisition of the three breast MRI protocols. The second strategy (S2) differed from S1 only by moving the table out of the scanner (with the participant still in the same position without movement) during the two-minute breaks. For the third strategy (S3) the participant got off the table during the two minutes breaks (Figure 1). In total, 18 different MRI exams were acquired for each healthy volunteer with a total scanning time of approximately 198 minutes.

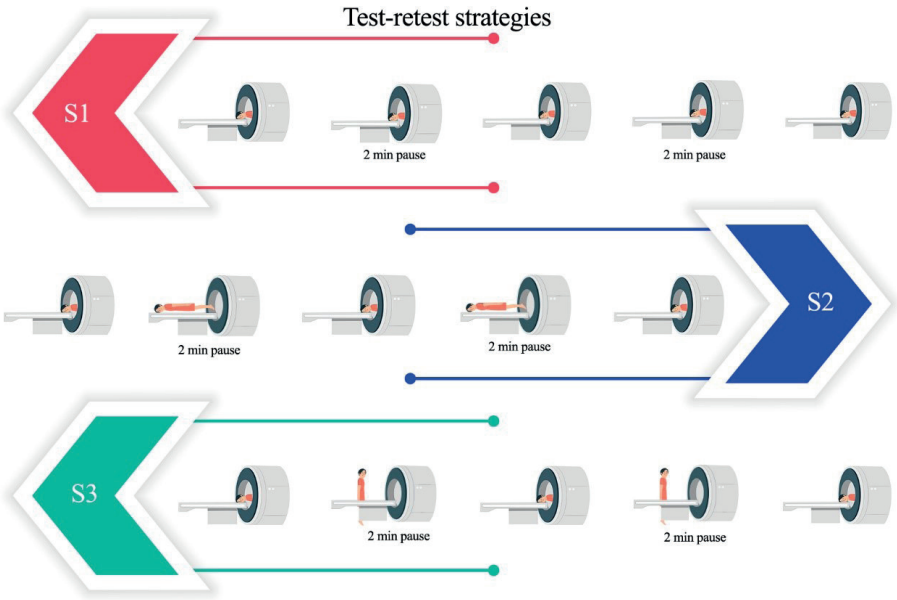


Figure 1. Visual representation of the three test-retest strategies

ROI segmentation

All images were visually checked for quality (including artifacts) by a dedicated breast radiologist with 14 years of experience (ML) before starting the analysis. The region of interest (ROI) was segmented by a medical researcher (RG) with four years of experience in breast MR imaging and validated by the same dedicated

breast radiologist. It was chosen to 3D, manually segment the right breast. The segmentations were bounded by the sternum (medial side), the pectoral muscle (dorsal side), and the axilla (lateral side) in three dimensions using MIM software (version 7.1.3, Cleveland Ohio, Unites States). Segmentations were performed on all patients on the T2W sequences of all MRI exams as anatomical structures are best visible on this sequence. Subsequently, the T2W sequence was registered with the T1W sequence, and ADC map, using rigid alignments within MIM software, followed by segmentations transfer (Figure 2).

Figure 2. An axial slice of a 3D MRI exam of a healthy volunteer including right breast segmentation (red margin). (A) ADC map, (B) T2-weighted image, (C) T1-weighted image

Image pre-processing & feature extraction.

All MRI exams including ROI segmentations were converted to the nearly raw raster data (NRRD) file format using Python (version 3.7.3) for subsequent analysis. Before feature extraction, multiple pre-processing procedures were applied to the images to study their impact on feature repeatability. First, feature extraction was performed without any image pre-processing as a baseline measurement. Second, N4 bias field correction was applied to the images prior to feature extraction [21]. Lastly, the bias field corrected images were further pre-processed using the built-in image z-score normalization by Pyradiomics software (version 2.2.0), with and without binning the voxel grayscale values using a fixed bin width of 32 and 64 (Pyradiomics suggested a bin width between 16-128) [16,22]. Image pre-processing steps were performed in Python (version 3.7) using an in-house developed pipeline based on the computer vision packages, including OpenCV (version 4.1.0), SimpleITK (version 1.2.0), and NumPy (version 1.16.2). For each ROI, 91 original features were extracted using the Pyradiomics software (version 3.0.1), which is mostly compliant with the Image Biomarker Standardization Initiative [23]. The extracted radiomics feature included first-order statistics features, gray-level co-occurrence matrix features (GLCM), gray-level run length matrix features (GLRLM), gray-level size zone matrix features (GLSZM), neighboring gray tone difference matrix features (NGTDM), and gray-level dependence matrix features (GLDM). All texture features were extracted using default Pyradiomics settings. A detailed Pyradiomics feature description can be found online [24].

Statistical analysis

To assess the repeatability of the extracted radiomic features for the various ROI's in the multiple test-retest strategies, the concordance correlation coefficient (CCC) was calculated using the epiR package (Version 0.9-99) (REF) in R language (version 3.6.3) performed in R studio (version 1.2.1335, Vienna Austria) [25]. Radiomics features extracted from a given MRI exam are compared to radiomic features extracted from the remaining MRI exams in a pairwise manner. The CCC was used to evaluate the agreement in radiomic feature values, taking into account both the rank and the value of the measurements [26]. This metric has the advantage of robust results in small sample sizes [26]. The CCC provides values between -1 and 1, with 0 representing no concordance, 1 representing perfect concordance, and -1 perfect inverse concordance. Features with a CCC of > 0.90 were defined as repeatable features, according to suggestions in literature [27]. Feature concordance was assessed for each pre-processing procedure using the results of all test-retest strategies of both scanning dates as well as for the results collected on the separate scanning dates. To create an overview of repeatable features across all pairs for the different pre-processing procedures, the intersection of the repeatable features across pairs was noted.

Results

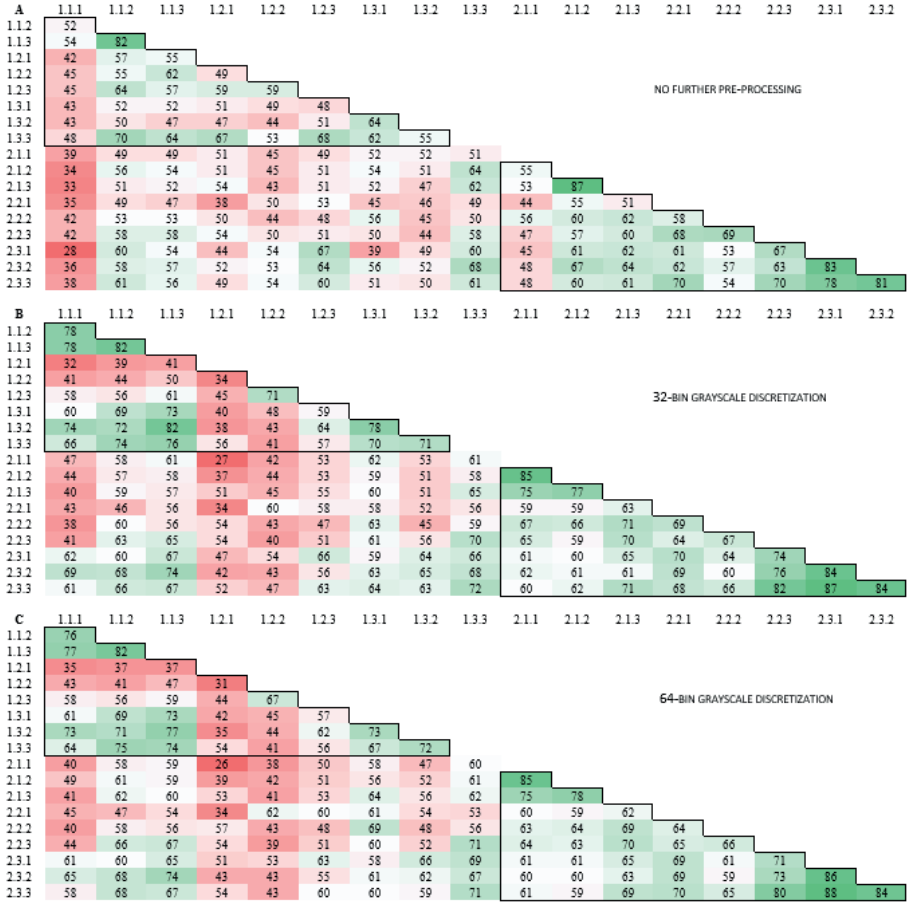
Patients demographics

The median age of the eleven healthy female volunteers was 28 years (interquartile range 25-30 years). Table 1 summarizes the healthy volunteers' characteristics. Shimbox displacement occurred in 22.6% of the scanned sequences.

Table 1. Patient characteristics

	healthy volunteers (n=11)
Age (years) (median; IQR)	28 (25 - 30)
Height (cm) (median; IQR)	167 (167 - 172)
Weigth (kg) (median; IQR)	60 (58 - 63)
Week of the menstrual cycle* Week 1 Week 2 Week 3 Week 4	Date 1 / Date 2 1 / 5 1 / 1 3 / 1 4 / 2
Days between scan (mean; range)	7 (6 - 9)

* no measurement of the menstrual cycle possible for two healthy volunteers


NO FURTHER PRE-PROCESSING
32-BIN GRAYSCALE DISCRETIZATION
64-BIN GRAYSCALE DISCRETIZATION

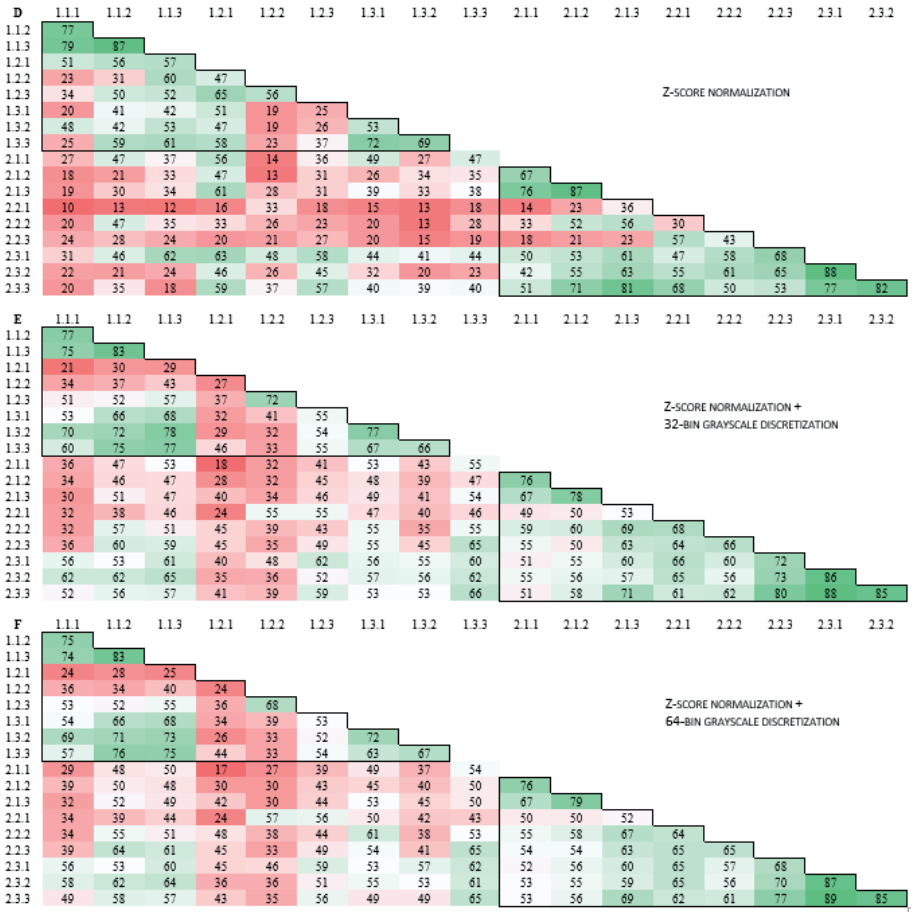


Figure 3. Number of pairwise concordant radiomic features using a concordance correlation coefficient > 0.90 for T1-weighted images with A: no further pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization. The black frame in the top left corner shows the MRI exams taken during the first scan date and the black frame in the bottom right corner shows the MRI exams taken during the second scan date. The numbers on the axis refer to the different MRI exams scanned, wherein the first number corresponds to the scan date and the second number to the test-retest strategy. In each test-retest strategy, three scans were examined which is represented by the last number. A total of 91 radiomic features was examined.

Repeatable radiomic features

Due to a scanning error of all T1-weighted images and the ADC maps of one healthy volunteer during scanning date 1, all data of this participant was excluded from the analysis. In both the T1W and T2W sequences as in the ADC maps, in pairwise comparison, the number of concordant features varied per scanning date, per test-retest strategy and, per image pre-processing procedure (Figure 3, 4, and 5). Furthermore, for all pre-processing procedures, the lowest number of concordant features was observed between the MRI exams scanned on date 1 and the MRI exams scanned on date 2, seen in the reddest field outside the black demarcations in Figures 3, 4 and 5.

T1W Sequence

Across all pairs, regardless of scanning date and test-retest strategy, the highest number of concordant features was seen in the images without pre-processing, resulting in 15/91 (16.5%) concordant features. These 15 features consisted of 7 first-order, 1 GLCM, 2 GLRLM, 2 GLSZM, and 2 GLDM and, 1 NGTDM feature(s) (Table 2). Applying grayscale discretization resulted in 13/91 (14.3%) and 14/91 (15.4%) concordant features for 32-bins and 64-bins, respectively. Compared to the images without pre-processing, the texture features showed less concordant features. The z-score normalized images resulted in the lowest number of 4/91 (4.4%) concordant features. Applying gray-scale discretization after z-score normalization improved the number of concordant textural features to 7/91 (7.7%) and 8/91 (8.8%) for 32-bins and 64-bins, respectively. The loss in the number of concordant features for z-score normalized images (with and without grayscale discretization), when compared to the images without pre-processing, was mainly due to a loss in the number of concordant first-order features, which were 6/91 (6.6%).

For the majority of pre-processing strategies, the images collected during date 2 showed a higher number of concordant features (varying between 10/91 and 48/91 in images without BFC and between 11/91 and 35/91 in BFC images) compared to images collected during date 1 (varying between 4/91 and 32/91 in images without BFC and between 9/91 and 14/91 in BFC images) (Table 3, Figure 3), with these differences being greatest after applying grayscale discretization. Furthermore, for most image pre-processing procedures, the addition of BFC resulted in less concordant features compared to the images without BFC (Table 3, Table S2). For the BFC images without further pre-processing and for the BFC images with grayscale discretization, it was mainly the first-order features that showed a loss of concordance compared to not performing BFC.

Table 2. Concordant features across all pairs for the T1-weighted MRI exams, with A: no pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization.

	A	B	C	D	E	D
Number of concordant features	**15 (16.5%)**	**13 (14.3%)**	**14 (15.4%)**	**4 (4.4%)**	**7 (7.7%)**	**8 (8.8%)**
firstorder_90Percentile	x	x	x			
firstorder_InterquartileRange	x	x	x			
firstorder_MeanAbsoluteDeviation	x	x	x			
firstorder_Mean	x	x	x			
firstorder_RobustMeanAbsoluteDeviation	x	x	x			
firstorder_RootMeanSquared	x	x	x			
firstorder_Skewness	x	x	x	x	x	x
glcm_JointAverage	x					
glrlm_GrayLevelNonUniformity	x	x	x		x	x
glrlm_RunLengthNonUniformity	x	x	x		x	x
glszm_GrayLevelNonUniformity	x		x	x		x
glszm_SizeZoneNonUniformity				x		
glszm_SmallAreaHighGrayLevelEmphasis	x					
gldm_DependenceNonUniformity	x	x	x		x	x
gldm_GrayLevelNonUniformity	x	x	x	x	x	x
ngtdm_Busyness		x	x		x	x
ngtdm_Coarseness	x	x	x		x	x

T2W sequence

Across all pairs, regardless of scanning date and test-retest strategy, the z-score normalized images showed the highest number of concordant features, 26/91 (28.6%), of which, 3 first-order, 11 GLCM, 3 GLRLM, 0 GLSZM, 8 GLDM, and 1 NGTDM feature(s) (Table 4). Compared to the other pre-processing procedures, the difference in the number of concordant features was mainly in the concordant texture features, which were almost non-concordant for the other pre-processing procedures.

The images without pre-processing resulted in 11/91 (12.1%) concordant features across all pairs, of which more than half of these features were first-order features (Table 4). Applying grayscale discretization resulted in a further decrease of concordant features to 7/91 (7.7%) for both 32 and 64 bins. Applying grayscale discretization after z-score normalization resulted in a loss of almost all concordant textural features when compared to z-score normalized images alone. These images resulted in only 4/91 (4.4%) concordant features for both 32 and 64 bins. Notably, the only concordant texture feature (gldm_SmallDependenceLowGrayLevelEmphasis) was not concordant after z-score normalization alone.

A

	1.1.1	1.1.2	1.1.3	1.2.1	1.2.2	1.2.3	1.3.1	1.3.2	1.3.3	2.1.1	2.1.2	2.1.3	2.2.1	2.2.2	2.2.3	2.3.1	2.3.2
1.1.2	52																
1.1.3	54	82															
1.2.1	42	57	55														
1.2.2	45	55	62	49													
1.2.3	45	64	57	59	59												
1.3.1	43	52	52	51	49	48											
1.3.2	43	50	47	47	44	51	64										
1.3.3	48	70	64	67	53	68	62	55									
2.1.1	39	49	49	51	45	49	52	52	51								
2.1.2	34	56	54	51	45	51	54	51	64	55							
2.1.3	33	51	52	54	43	51	52	47	62	53	87						
2.2.1	35	49	47	38	50	53	45	46	49	44	55	51					
2.2.2	42	53	53	50	44	48	56	45	50	56	60	62	58				
2.2.3	42	58	58	54	50	51	50	44	58	47	57	60	68	69			
2.3.1	28	60	54	44	54	67	39	49	60	45	61	62	61	53	67		
2.3.2	36	58	57	52	53	64	56	52	68	48	67	64	62	57	63	83	
2.3.3	38	61	56	49	54	60	51	50	61	48	60	61	70	54	70	78	81

NO FURTHER PRE-PROCESSING

B

	1.1.1	1.1.2	1.1.3	1.2.1	1.2.2	1.2.3	1.3.1	1.3.2	1.3.3	2.1.1	2.1.2	2.1.3	2.2.1	2.2.2	2.2.3	2.3.1	2.3.2
1.1.2	78																
1.1.3	78	82															
1.2.1	32	39	41														
1.2.2	41	44	50	34													
1.2.3	58	56	61	45	71												
1.3.1	60	69	73	40	48	59											
1.3.2	74	72	82	38	43	64	78										
1.3.3	66	74	76	56	41	57	70	71									
2.1.1	47	58	61	27	42	53	62	53	61								
2.1.2	44	57	58	37	44	53	59	51	58	85							
2.1.3	40	59	57	51	45	55	60	51	65	75	77						
2.2.1	43	46	56	34	60	58	58	52	56	59	59	63					
2.2.2	38	60	56	54	43	47	63	45	59	67	66	71	69				
2.2.3	41	63	65	54	40	51	61	56	70	65	59	70	64	67			
2.3.1	62	60	67	47	54	66	59	64	66	61	60	65	70	64	74		
2.3.2	69	68	74	42	43	56	63	65	68	62	61	61	69	60	76	84	
2.3.3	61	66	67	52	47	63	64	63	72	60	62	71	68	66	82	87	84

32-BIN GRAYSCALE DISCRETIZATION

C

	1.1.1	1.1.2	1.1.3	1.2.1	1.2.2	1.2.3	1.3.1	1.3.2	1.3.3	2.1.1	2.1.2	2.1.3	2.2.1	2.2.2	2.2.3	2.3.1	2.3.2
1.1.2	76																
1.1.3	77	82															
1.2.1	35	37	37														
1.2.2	43	41	47	31													
1.2.3	58	56	59	44	67												
1.3.1	61	69	73	42	45	57											
1.3.2	73	71	77	35	44	62	73										
1.3.3	64	75	74	54	41	56	67	72									
2.1.1	40	58	59	26	38	50	58	47	60								
2.1.2	49	61	59	39	42	51	56	52	61	85							
2.1.3	41	62	60	53	41	53	64	56	62	75	78						
2.2.1	45	47	54	34	62	60	61	54	53	60	59	62					
2.2.2	40	58	56	57	43	48	69	48	56	63	64	69	64				
2.2.3	44	66	67	54	39	51	60	52	71	64	63	70	65	66			
2.3.1	61	60	65	51	53	63	58	66	69	61	61	65	69	61	71		
2.3.2	65	68	74	43	43	55	61	62	67	60	60	63	69	59	73	86	
2.3.3	58	68	67	54	43	60	60	59	71	61	59	69	70	65	80	88	84

64-BIN GRAYSCALE DISCRETIZATION

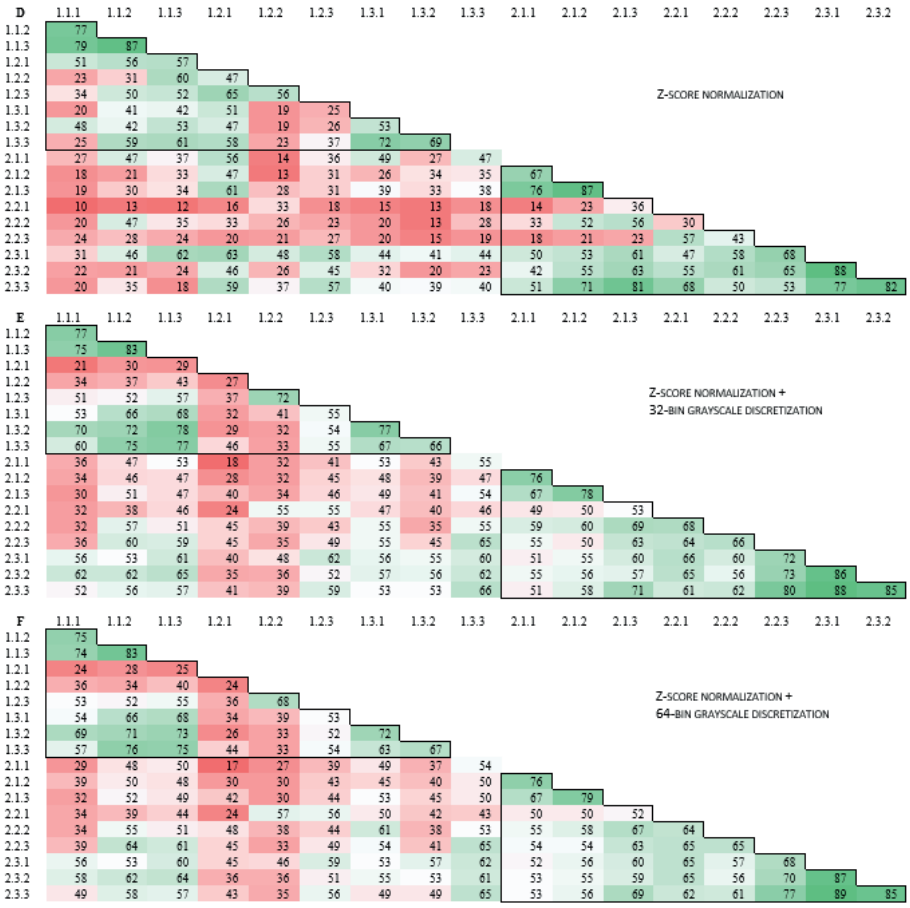


Figure 4. Number of pairwise concordant radiomic features using a concordance correlation coefficient > 0.90 for T2-weighted images with A: no further pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization. The black frame in the top left corner shows the MRI exams taken during the first scan date and the black frame in the bottom right corner shows the MRI exams taken during the second scan date. The numbers on the axis refer to the different MRI exams scanned, wherein the first number corresponds to the scan date and the second number to the test-retest strategy. In each test-retest strategy, three scans were examined which is represented by the last number. A total of 91 radiomic features was examined.

Table 3. Number of concordant features across all pairs for the entire dataset (All) and across all pairs from the separate scanning dates (Date 1 and Date 2) for all sequences with and without bias field correction (BFC), with A: no further pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization

Sequences	Without BFC			With BFC		
	All	Date 1	Date 2	All	Date 1	Date 2
T1W						
A	15	32	40	8	13	11
B	13	19	45	10	11	30
C	14	18	48	8	12	31
D	4	4	10	4	9	12
E	7	10	35	10	13	34
F	8	9	38	8	14	35
T2W						
A	11	31	16	0	1	60
B	7	9	12	2	3	22
C	7	9	11	1	3	23
D	26	35	44	26	39	37
E	4	7	7	6	11	17
F	4	7	6	5	11	18
ADC						
A	8	28	22	8	9	12
B	7	15	13	6	9	12
C	6	11	11	6	11	11

The addition of BFC resulted in different feature concordance when compared to the same image pre-processing procedures without BFC (Table 4, Table S3). The BFC images without further pre-processing, with 32-bin grayscale discretization and, with 64-bin grayscale discretization resulted in 0/91 (0.0%), 2/91 (2.2%), and 1/91 (1.1%) concordant features, respectively. Despite the overall loss of concordant features, 2/91 (2.2%) features were found to be concordant after the addition of BFC. The BFC z-score normalized images showed the same number of concordant features compared to the z-score normalized images without BFC, although some features improved in concordance, where others lost concordance. The application of grayscale discretization after z-score normalization on BFC images showed the same pattern in concordant features when compared to the images without BFC, namely, a loss of almost all concordant textural features (Table 4 and S3). These pre-processing procedures resulted in 6/91 (6.6%) and 5/91 (5.5%) concordant features, for 32-bins and 64-bins, respectively. Furthermore, it is noteworthy that when looking at the pairwise concordance features for the different scan dates, BFC decreased the feature concordance for MRI exams scanned on date 1, while there was an increase in feature concordance for MRI exams scanned on date 2 (Figure 4, Table 3).

Table 4. Concordant features across all pairs for the T2-weighted MRI exams, with A: no pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization

	A	B	C	D	E	F
Number of concordant features	**11 (12.1%)**	**7 (7.7%)**	**7 (7.7%)**	**26 (28.6%)**	**4 (4.4%)**	**4 (4.4%)**
firstorder_10Percentile				x	x	x
firstorder_90Percentile	x	x	x			
firstorder_InterquartileRange	x	x	x	x	x	x
firstorder_MeanAbsoluteDeviation	x	x	x			
firstorder_Mean	x	x	x			
firstorder_RobustMeanAbsoluteDeviation	x	x	x	x	x	x
firstorder_RootMeanSquared	x	x	x			
glcm_JointAverage	x					
glcm_Contrast				x		
glcm_DifferenceAverage	x			x		
glcm_DifferenceEntropy				x		
glcm_DifferenceVariance				x		
glcm_JointEntropy				x		
glcm_Idm				x		
glcm_Idmn				x		
glcm_Id				x		
glcm_Idn				x		
glcm_InverseVariance				x		
glcm_SumEntropy				x		
glrlm_GrayLevelNonUniformity				x		
glrlm_RunLengthNonUniformity	x					
glrlm_RunPercentage				x		
glrlm_RunVariance				x		
gldm_DependenceEntropy				x		
gldm_DependenceNonUniformity				x		
gldm_ DependenceNonUniformityNormalized				x		
gldm_DependenceVariance				x		
gldm_GrayLevelNonUniformity				x		
gldm_LargeDependenceEmphasis				x		
gldm_ LargeDependenceHighGrayLevelEmphasis				x		
gldm_ SmallDependenceHighGrayLevelEmphasis				x		
gldm_ SmallDependenceLowGrayLevelEmphasis	x	x	x		x	x
ngtdm_Complexity				x		
ngtdm_Contrast	x					

ADC map

Across all pairs, regardless of scanning date and test-retest strategy, the number of concordant features for the images without pre-processing, with 32-bin grayscale discretization, and 64-bin grayscale discretization was 8/91 (8.8%), 7/91 (7.7%), and 6 (6.6%), respectively (Table 5). In none of the pre-processing procedures, first-order features appeared to be concordant. The number of concordant features was roughly the same for the BFC images with 8/91 (8.8%), 6/91 (6.6%), and 6/91 (6.6%) concordant features for images without further pre-processing, with 32-bin grayscale discretization, and 64-bin grayscale discretization, respectively (Table 5). Although compared to the images without BFC, some features improved in concordance, where others lost concordance (Table 5).

The number of concordant features differed between the images collected on the separated scanning dates, although these differences were minor compared to the T1W and T2W sequences (Figure 5, Table 3). The number of concordant features was 28/91 (30.8%), 15/91 (16.5%) and 11/91 (12.1%) for date 1 and 22/91 (24.1%), 13/91 (14.3%) and 11/91 (12.1%) for date 2, using the images without BFC. The number of concordant features was 9/91 (9.9%), 9/91 (9.9%) and 11/91 (12.1%) for date 1 and 12/91 (13.2%), 12/91 (13.2%) and 11/91 (12.1%) for date 2, using the BFC images.

Table 5. Concordant features across all pairs for the ADC maps, with A: no pre-processing, B: 32-bin grayscale discretization, and C: 64-bin grayscale discretization, D: bias field correction, E: bias field correction + 32-bin grayscale discretization and, F: bias field correction + 64-bin grayscale discretization.

	A	B	C	D	E	F
Number of concordant features	**8 (8.8%)**	**7 (7.7%)**	**6 (6.6%)**	**8 (8.8%)**	**6 (6.6%)**	**6 (6.6%)**
glcm_ClusterProminence	x					
glcm_Correlation	x	x	x	x	x	x
glcm_Imc1	x	x	x	x		x
glcm_Imc2	x	x	x	x	x	x
glrlm_GrayLevelNonUniformity		x	x		x	x
glrlm_RunLengthNonUniformity	x	x	x	x	x	x
glszm_GrayLevelNonUniformity	x	x	x	x	x	x
glszm_SizeZoneNonUniformity	x			x		
gldm_DependenceNonUniformity	x			x		
ngtdm_Coarseness		x		x	x	

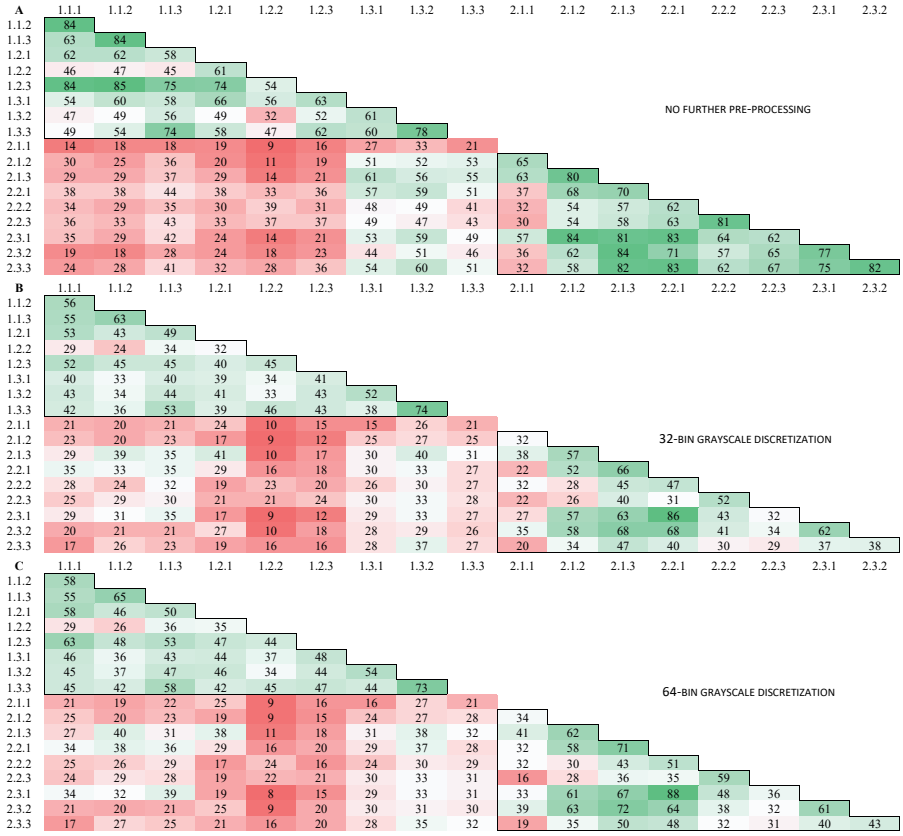


Figure 5. Number of pairwise concordant radiomic features using a concordance correlation coefficient > 0.90 for ADC maps with A: no further pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization. The black frame in the top left corner shows the MRI exams taken during the first scan date and the black frame in the bottom right corner shows the MRI exams taken during the second scan date. The numbers on the axis refer to the different MRI exams scanned, wherein the first number corresponds to the scan date and the second number to the test-retest strategy. In each test-retest strategy, three scans were examined which is represented by the last number. A total of 91 radiomic features was examined.

Discussion

In this test-retest study, repeatable radiomics features extracted from breast MRI exams from healthy volunteers were identified, using a fixed scanning protocol including T2-weighted (T2W), unenhanced T1-weighted (T1W), and diffusion-weighted images with corresponding derived ADC maps. This study showed the effects of varying image pre-processing procedures on the radiomics feature repeatability. Across all pairs, the images without pre-processing produced the highest number of repeatable features for both the T1W sequence as well as the ADC maps. In the T2W images, applying z-score normalization produced the highest number of repeatable features.

The assessment of radiomics feature repeatability via test-retest studies in breast MRI exams is currently lacking. The three different MRI sequences examined in this study showed differences in feature repeatability. In addition, the effect of image pre-processing on feature repeatability was different for the two MRI sequences and ADC maps. Not applying image pre-processing produced the highest number of repeatable features in the T1W sequence and the ADC maps. Overall, applying grayscale discretization caused a loss of repeatable textural features in the T1W and T2W sequences, although some texture features became repeatable after grayscale discretization. It is notable that in general, the number of repeatable texture features was reduced after applying grayscale discretization, although grayscale discretization is considered necessary for the extraction of texture features by both Pyradiomics and the IBSI guidelines [22]. Given that MR images do not contain absolute signal values, MRI exams performed on the same scanner using an identical scan protocol could potentially eliminate the need for grayscale discretization. Furthermore, z-score normalized images showed the highest number of repeatable features in the T2W sequence, on the other hand, applying normalization decreased the number of repeatable features in the T1W sequence. Failure to improve the repeatability of features after z-score normalization was also found in the study by Schwier et al. [19], although, in contrast to our results, this was seen in the T2W sequence. They state that image normalization was used to homogenize images acquired from different scanners with different protocols. In our study, however, it was assumed that images scanned with the same protocol on the same scanner were already well comparable in terms of imaging parameters. In addition, the applied normalization uses the whole image for normalization and since the MRI quality decreases further from the coil (at the edges of the images), this reduction in quality can degrade the quality of the breast region (which is close to the coil) and with that the ROI comparability. The same principle could account for the use of BFC since for all sequences it either did not change the number of repeatable features or caused a loss of repeatable features compared to not using BFC. However, failure to improve the repeatability

of functions after BFC may also be due to use of default settings for the N4 BFC; findings of Saint Martin et al. [28] showed that the default settings for the N4 BFC were not optimal for breast MRI exams.

By considering pairwise comparisons between scans taken on the same day, it was found that for all sequences, including all different preprocessing procedures, except for the T2W sequence and ADC maps without preprocessing, date 2 produced a higher number of repeatable features compared to date 1. One explanation for this may be that the healthy volunteers knew better what to expect on the 2nd scan date after going through the first scan date. In addition, in most cases, the number of repeatable features was higher for the scans taken on the same day compared to the number of repeatable features found from the data of both days, as expected. These differences may be explained by changing factors over time (e.g., system changes in the MRI scanner or biology of the healthy volunteer) that caused variation in the feature values. For example, the homogeneity of the MRI field, gradient systems, and coil affects the image quality[29]. Furthermore, changes in the biology of the healthy volunteer, including the menstrual cycle and body temperature, are known to affect the MRI exams [30]. These factors may impact clinical decision making and hence, radiomic features must be robust to these changes.

To date, MRI test-retest studies for the evaluation of repeatable and reproducible features, have been conducted through phantom research [15,28,31-33] and by the use of MRI exams of healthy volunteers or cancer patients [17,19,20,32,34-36]. None of these studies investigated feature repeatability and/or reproducibility in human breast MRI exams, and only one study investigated a breast phantom [28]. The study of Saint Martin et al. [28] showed the necessity of image pre-processing dedicated to breast MRI exams before using features in further analysis. Phantom repeatability and reproducibility results seem to be overly optimistic as these overall appear to score higher than the test-retest studies performed within human data. For example, the study by Lee et al. [32] tested feature repeatability in T1W and T2W in both a phantom and MRI brain of healthy volunteers. The average ICC repeatability measures for the T1W and T2W images were higher for the phantom (0.963 and 0.959) compared to healthy volunteers (0.856 and 0.849). Furthermore, a recently published phantom study by Shur et al. [31] showed that 37/46 (80%) of the radiomic features were concordant (CCC > 0.9) in a test-retest study. By contrast, the test-retest study by Eck et al. [34] investigating feature repeatability in T2W brain MRI exams of fifteen healthy volunteers showed only 76/146 (52%) of good to excellent repeatable features (CCC ≥ 0.7). Considering only the excellent repeatable features (CCC > 0.85) in the above-mentioned article, the number of concordant features decreased to 40/146 (27.4%), which is more comparable to the results found in this study. The same accounts for a test-retest study in brain

MRI exams of glioblastoma patients, in which they identified 386/1043 (37.0%) repeatable features, although they used CCC > 0.8 as a cut-off value [36]. A prostate MRI repeatability study by Schwier et al. [19] concluded that feature repeatability can vary greatly among the radiomic features and that the repeatability of the features is highly sensitive to image pre-processing procedures.

In clinical (prospective) trials, variance in scanners and acquisition and reconstruction parameters between and even within patients is unsurmountable and will therefore affect the reproducibility of the features. Although exploring feature reproducibility was not the aim of this study, this data will be a starting point to investigate the reproducibility of breast MRI extracted radiomic features. Future studies can investigate feature reproducibility by changing the different acquisition parameters one by one while leaving the others fixed. Furthermore, the harmonization method called ComBat, which was originally developed to harmonize gene expression data [37], is increasingly being applied in radiomics studies to remove batch effects [8,14,38-40]. However, caution should be exercised when applying this harmonization method, as it can only correct for one variable and, MRI data collected from multiple hospitals often contains a multitude of variables. In addition, future studies should focus on the discriminative power of a repeatable and reproducible feature, as a repeatable and reproducible feature does not necessarily imply that this feature is a predictive or prognostic radiomic feature.

Limitations

Firstly, the number of healthy volunteers included was quite limited, although the test-retest set-up allowed for 18 MRI exams per healthy volunteer, resulting in the analysis of a total of 198 MRI exams. Nevertheless, since this is an early study investigating this topic, we believe that these results are valuable and useful for the radiomics community. Secondly, the included T1W images were examined without adding a contrast agent, so these images cannot be fully compared to the dynamic T1W images normally examined in a clinical breast protocol. Future test-retest studies in breast cancer patients should show whether the repeatable features found in this study are also repeatable in dynamic T1W images. Thirdly, this study investigated only Pyradiomics features extracted from the original image. Future studies could focus more on other feature groups, among others, Gabor, gradient, or Laws. Fourthly, the region of interest contained only healthy tissue, further research in breast cancer patients will have to show whether the repeatable features found in healthy breast tissue can also be considered repeatable in breast tumor tissue. Lastly, it is important to keep in mind that there is a great variety of pre-processing procedures, which can influence feature values. In this study, we choose to use the open-source software Pyradiomics to apply normalization and grayscale discretization to easily reproduce results. In the future, we aim to extend this study with other alternative normalization procedures and focus on feature repeatability.

Conclusion

Varying numbers of repeatable breast MRI radiomic features extracted from healthy volunteers were found for each different test-retest strategy. Furthermore, the effects of image preprocessing procedures on the repeatability of radiomic features were found to be different depending on the MRI sequence.

Supplementary Materials

Table S1. Acquisition parameters

	RT	TE	ST	FA	WFS	ETL	PS	AM	NoS
T2W	2000	223	2	90	0.28	95	0.79 x 0,79	340, 339	220
T1W	5.30	3.0	2	10	0.39	38	0.36 x 0.36	453, 450	170
DWI	10765	88	3	90	9.46	61	1.01 x 1.01	151, 146	150

Abbreviations: RT, repetition time; TE, echo time; ST, slice thickness; FA, flip angle; WFS, water-fat shift; ETL, echo train-length; PS, pixel spacing; AM, acquisition matrix; NoS, number of slices; T2W, T2-weighted; T1W, T1-weighted; DWI, diffusion-weighted image.

Table S2. Concordant features across all pairs for the bias field corrected T1-weighted MRI exams, with A: no further pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization

	A	B	C	D	E	F
Number of concordant features	**8 (8.8%)**	**10 (11.0%)**	**8 (8.8%)**	**4 (4.4%)**	**10 (11.0%)**	**8 (8.8%)**
firstorder_Skewness	x	x	x	x	x	x
firstorder_Uniformity		x	x		x	x
glrlm_GrayLevelNonUniformity	x	x	x		x	x
glrlm_GrayLevelNonUniformityNormalized		x	x		x	x
glrlm_RunLengthNonUniformity	x		x			x
glszm_GrayLevelNonUniformity	x		x	x		
glszm_LargeAreaHighGrayLevelEmphasis		x			x	
glszm_SizeZoneNonUniformity	x			x		
gldm_DependenceEntropy		x			x	
gldm_DependenceNonUniformity	x	x			x	
gldm_GrayLevelNonUniformity	x	x		x	x	x
ngtdm_Busyness		x	x		x	x
ngtdm_Coarseness	x	x	x		x	x

Table S3. Concordant features across all pairs for the bias field corrected T2-weighted MRI exams, with A: no further pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization.

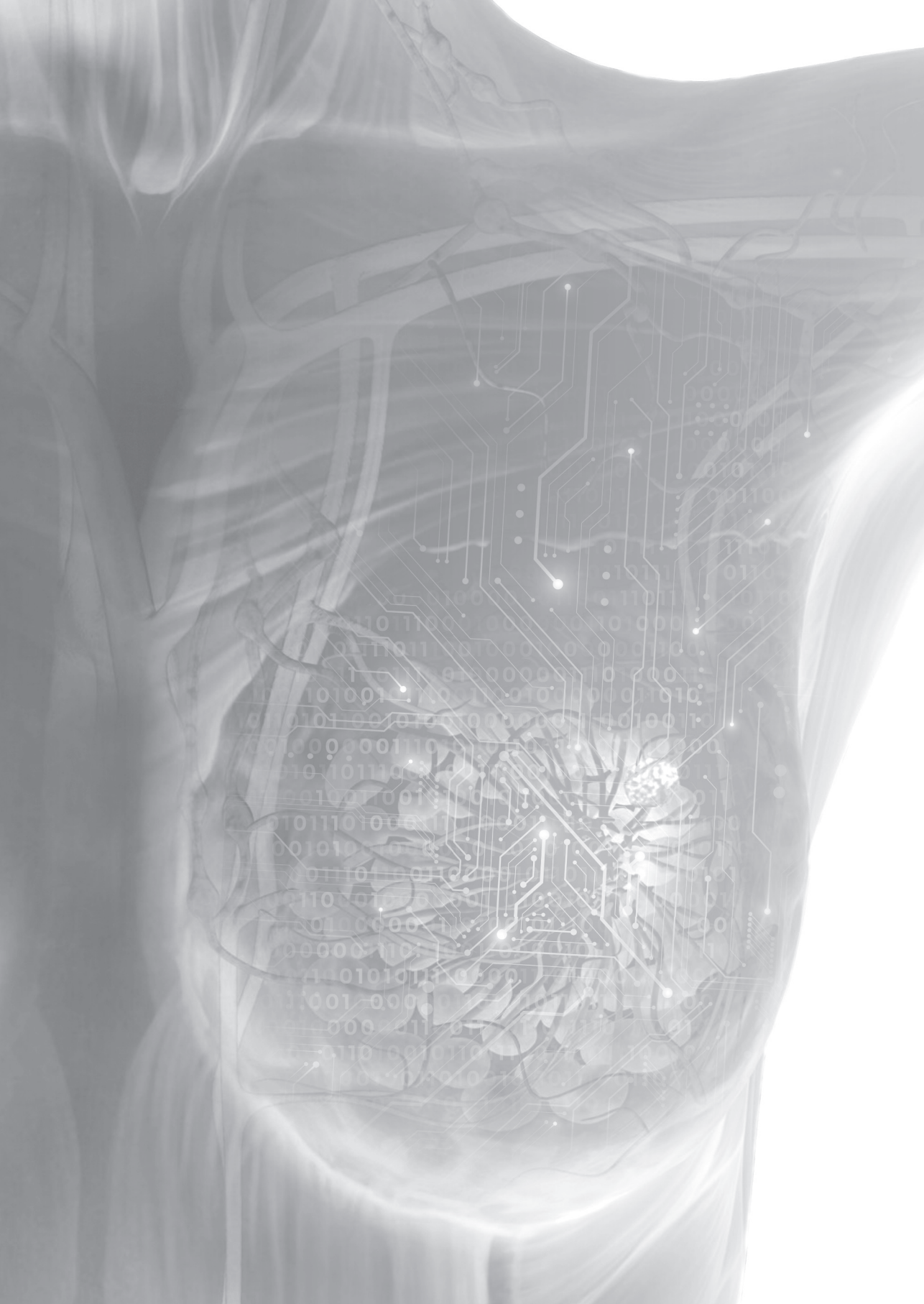
	A	B	C	D	E	D
Number of stable features	**0 (0.0%)**	**2 (2.2%)**	**1 (1.1%)**	**26 (28.6%)**	**6 (6.6%)**	**5 (5.5%)**
firstorder_10Percentile				x	x	x
firstorder_InterquartileRange				x	x	x
firstorder_Kurtosis			x			
firstorder_MeanAbsoluteDeviation				x	x	x
firstorder_RobustMeanAbsoluteDeviation				x	x	x
glcm_Contrast				x		
glcm_DifferenceAverage				x		
glcm_DifferenceEntropy				x		
glcm_DifferenceVariance				x		
glcm_JointEntropy				x		
glcm_Idm				x		
glcm_Idmn				x		
glcm_Id				x		
glcm_Idn				x		
glcm_InverseVariance				x		
glcm_SumEntropy				x		
glrlm_GrayLevelNonUniformity		x	x		x	x
glrlm_RunPercentage				x		
glrlm_RunVariance				x		
gldm_DependenceEntropy				x		
gldm_DependenceNonUniformity				x		
gldm_DependenceNonUniformityNormalized				x		
gldm_DependenceVariance				x		
gldm_GrayLevelNonUniformity				x		
gldm_LargeDependenceEmphasis				x		
gldm_SmallDependenceEmphasis				x		
gldm_SmallDependenceHighGrayLevelEmphasis				x		
gldm_SmallDependenceLowGrayLevelEmphasis		x			x	
ngtdm_Complexity				x		

References

1. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749-762.
2. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006.
3. Ibrahim A, Vallières M, Woodruff H, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Seminars in Nuclear Medicine.* 2019.
4. Liu Z, Li Z, Qu J, et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res.* 2019.
5. Whitney HM, Taylor NS, Drukker K, et al. Additive Benefit of Radiomics Over Size Alone in the Distinction Between Benign Lesions and Luminal A Cancers on a Large Clinical Breast MRI Dataset. *Acad Radiol.* 2019;26(2):202-209.
6. Bickelhaupt S, Paech D, Kickingereder P, et al. Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography. *J Magn Reson Imaging.* 2017;46(2):604-616.
7. Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N. Radiomics in breast cancer classification and prediction. *Semin Cancer Biol.* 2020.
8. Ibrahim A, Primakov S, Beuque M, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods.* 2020.
9. Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N. Overview of radiomics in breast cancer diagnosis and prognostication. *The Breast.* 2020;49:74-80.
10. Granzier RWY, van Nijnatten TJA, Woodruff HC, Smidt ML, Lobbes MBI. Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review. *European journal of radiology.* 2019;121:108736.
11. Simpson G, Ford JC, Llorente R, et al. Impact of quantization algorithm and number of gray level intensities on variability and repeatability of low field strength magnetic resonance image-based radiomics texture features. *Phys Med.* 2020;80:209-220.
12. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res.* 2015;24(1):27-67.
13. Varghese BA, Hwang D, Cen SY, et al. Reliability of CT-based texture features: Phantom study. *Journal of Applied Clinical Medical Physics.* 2019;20(8):155-163.
14. Ibrahim A, Refaee T, Primakov S, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers.* 2021;13(8).
15. Baessler B, Weiss K, Pinto Dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest Radiol.* 2019;54(4):221-228.
16. Hoebel KV, Patel JB, Beers AL, et al. Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma. *Radiology: Artificial Intelligence.* 2021;3(1).
17. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2019;135:107-114.

18. Fedorov A, Vangel MG, Tempany CM, Fennessy FM. Multiparametric Magnetic Resonance Imaging of the Prostate: Repeatability of Volume and Apparent Diffusion Coefficient Quantification. *Invest Radiol.* 2017;52(9):538-546.
19. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci Rep.* 2019;9(1):9441.
20. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep.* 2019;9(1):4800.
21. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging.* 2010;29(6):1310-1320.
22. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017;77(21):e104-e107.
23. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology.* 2020.
24. Pyradiomics feature description. https://pyradiomics.readthedocs.io/en/latest/features.html.
25. Racine JS. RStudio: a platform-independent IDE for R and Sweave. *Journal of Applied Econometrics.* 2012;27(1):167-172.
26. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45(1):255-268.
27. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client Report: HAM2005-062.* 2005.
28. Saint MJS, Orlhac F, Akl P, et al. A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study. *Magn Reson Mater Phy.* 2021;34(3):355-366.
29. Jackson E. MR Acceptance Testing and Quality Control: Report of AAPM MR Subcommittee TG1. *Med Phys Journal Translated Name Medical Physics.* 2009;36(6).
30. Dontchos BN, Rahbar H, Partridge SC, Lehman CD, DeMartini WB. Influence of Menstrual Cycle Timing on Screening Breast MRI Background Parenchymal Enhancement and Diagnostic Performance in Premenopausal Women. *J Breast Imaging.* 2019;1(3):205-211.
31. Shur J, Blackledge M, D'Arcy J, et al. MRI texture feature repeatability and image acquisition factor robustness, a phantom study and in silico study. *Eur Radiol Exp.* 2021;5(1):2.
32. Lee J, Steinmann A, Ding Y, et al. Radiomics feature robustness as measured using an MRI phantom. *Sci Rep.* 2021;11(1):3973.
33. Dreher C, Kuder TA, Konig F, et al. Radiomics in diffusion data: a test-retest, inter- and intra-reader DWI phantom study. *Clin Radiol.* 2020;75(10):798 e713-798 e722.
34. Eck B, Chirra PV, Muchhala A, et al. Prospective Evaluation of Repeatability and Robustness of Radiomic Descriptors in Healthy Brain Tissue Regions In Vivo Across Systematic Variations in T2-Weighted Magnetic Resonance Imaging Acquisition Parameters. *J Magn Reson Imaging.* 2021.
35. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *J Appl Clin Med Phys.* 2020;21(1):179-190.
36. Kickingereder P, Neuberger U, Bonekamp D, et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro Oncol.* 2018;20(6):848-857.

37. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-127.
38. Fortin J-P, Parker D, Tunç B, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage.* 2017;161:149-170.
39. Fortin J-P, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage.* 2018;167:104-120.
40. Ibrahim A, Refaee T, Leijenaar RTH, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS One.* 2021;16(5):e0251147.



CHAPTER 7

MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability

Renée W. Y. Granzier, Nienke M. H. Verbakel[*], Abdalla Ibrahim[*], Janita E. van Timmeren, Thiemo J. A. van Nijnatten, Ralph T. H. Leijenaar RTH, Marc B. I. Lobbes, Marjolein L. Smidt[†], Henry. C. Woodruff[†]

[*] *Shared authorship*

[†] *Shared last authorship*

Sci Rep. 2020 Aug 25;10(1):14163

Abstract

Radiomics is an emerging field using the extraction of quantitative features from medical images for tissue characterization. While MRI-based radiomics is still at an early stage, it showed some promising results in studies focusing on breast cancer patients in improving diagnoses and therapy response assessment. Nevertheless, the use of radiomics raises a number of issues regarding feature quantification and robustness. Therefore, our study aim was to determine the robustness of radiomics features extracted by two commonly used radiomics software with respect to variability in manual breast tumor segmentation on MRI. A total of 129 histologically confirmed breast tumors were segmented manually in three dimensions on the first post-contrast T1-weighted MR exam by four observers: a dedicated breast radiologist, a resident, a Ph.D. candidate, and a medical student. Robust features were assessed using the intraclass correlation coefficient (ICC $>$0.9). The inter-observer variability was evaluated by the volumetric Dice Similarity Coefficient (DSC). The mean DSC for all tumors was 0.81 (range 0.19-0.96), indicating a good spatial overlap of the segmentations based on observers of varying expertise. In total, 41.6% (552/1328) and 32.8% (273/833) of all RadiomiX and Pyradiomics features, respectively, were identified as robust and were independent of inter-observer manual segmentation variability.

Introduction

Radiomics is a technique that is used to extract large amounts of quantitative information from routine medical images that decode information about a region of interest (ROI). The majority of radiomics articles published concerns its application in the oncological field [1-4]. Here, radiomics bears the advantage of non-invasively quantifying the underlying phenotype of the entire tumor for multiple lesions simultaneously, in contrast to tissue biopsy, which samples only a small part of a single (often heterogeneous) tumor [2,5]. The ability to characterize the tumor and to establish links to the underlying biology [6] and ultimately clinical outcomes, allows a more patient-tailored treatment [7], enabling 'precision medicine' [8,9]. Recently, several articles have outlined the potential clinical applicability of radiomics in the field of breast cancer for different purposes, *e.g.* diagnosis [10,11], tumor response prediction [12-14], prediction of molecular tumor subtype [15,16], and prediction of axillary lymph node metastases [17,18].

Although these results are promising, issues regarding features robustness as well as the comparability of results, including inter-observer segmentation variability, need to be addressed [19-24]. In order to extract clinically useful information from medical images and to use features as clinical biomarkers, it is important that extracted features are reproducible, standardized and robust [25,26]. All consecutive steps in the radiomics workflow induce potential uncertainties regarding feature robustness [27,28]. Since there used to be no gold standard or guideline for extraction of image features for radiomics use, an initiative –Image Biomarker Standardization Initiative (IBSI)- was launched as an effort to standardize the entire radiomics extraction process and encourage feature robustness [29].

ROI segmentation is an important step after image acquisition in the radiomics workflow, and one of the largest bottlenecks [30]. Traditionally, the edges (2D) or surfaces (3D) of the ROI are segmented, thereby defining a region from which features will be extracted. Segmentation can be performed either manually, semi-automatically, or completely automatically. Both manual and semi-automatic segmentation are prone to inter- and intra-observer variabilities, with the degree of observer experience playing an important role [31-33].

To the best of our knowledge, no articles have been published on the effect of manual inter-observer segmentation variability on MRI-based feature robustness in breast cancer patients. MRI is the most accurate modality for neoadjuvant systemic therapy response monitoring in breast cancer patients and as such much used in daily clinical practice [34-37]. In this article, we investigate the robustness of MR radiomics features, extracted using two commonly used radiomics software, with respect to variations in manual tumor segmentation of breast cancer patients.

Material and methods

Study population

In this single-center retrospective study, we collected data on 138 patients with histologically confirmed invasive breast cancer, who were planned for receiving NST and underwent a pretreatment DCE-MRI between January 2011 and December 2017 in Maastricht University Medical Center+. The institutional research board of the MUMC+ approved the study and waived the requirement for informed consent and the further need of guidelines. Exclusion criteria were: pathologically confirmed mastitis carcinomatosa, MR scan artifacts, or refusal of medical record usage by the patient. Furthermore, we excluded patients that underwent breast MRI exams with non-standard acquisition parameters, due to the use of a different MR scanner. All histologically confirmed breast tumors were included in the analysis. The complete process is summarized in the flowchart presented in Figure 1.

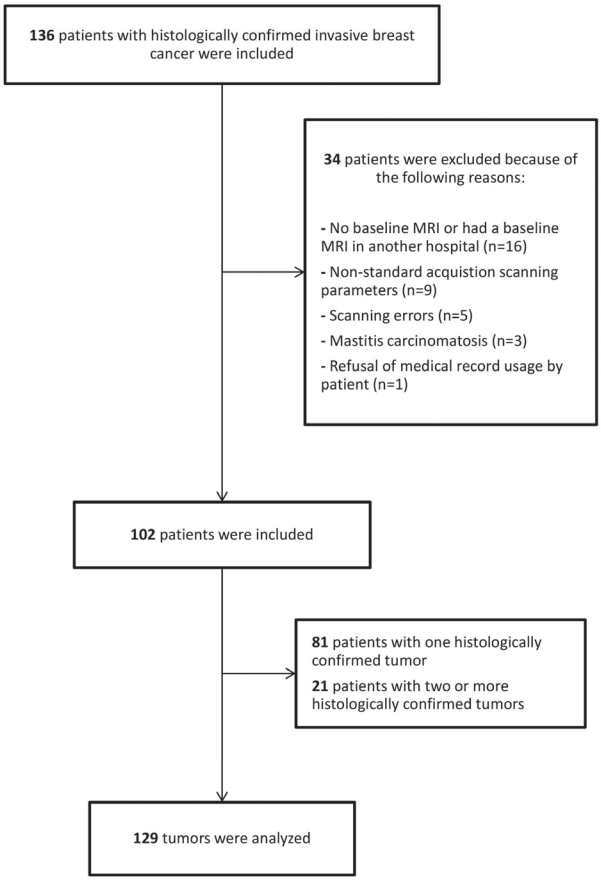


Figure 1. Flowchart of the patient population in the study.

Imaging data

All images were acquired by two clinically interchangeable (i.e. provide qualitatively similar images) 1.5T MRI scanners (Philips Intera and Philips Ingenia), using a dynamic contrast-enhanced T1-weighted (DCE-T1W) sequence with similar acquisition protocols (Table 1). The patients were scanned in prone position using a 16-channel dedicated breast coil. The DCE-T1W images were obtained before and after intravenous injection of gadolinium-based contrast Gadobutrol (Gadovist (EU)) with a volume of 15 cc and a flow rate of 2 ml/sec. One pre-contrast image and five post-contrast images were obtained for each patient.

Table 1. Imaging parameters for the breast DCE T1W sequence for both scanners.

	Scanner 1 Philips Ingenia (n)	**Scanner 2 Philips Intera (n)**
Number of tumors	100	29
Field strength (T)	1.5	1.5
Slice thickness (mm)	1.0	1.0
Repetition time (msec)	7.5 (88), 7.6 (12)	7.4 (13), 7.5 (15), 7.6 (1)
Echo time (msec)	3.4	3.4
Flip angle (degrees)	10	10
Echo train length	89* (range 62-175)	80* (range 60 – 85)
Pixel spacing (mm)	0.792 (3), 0.852 (1), 0.922 (2), 0.952 (47), 0.952 (47)	0.852 (1), 0.942 (1), 0.972 (26), 0.992 (1)
Temporal resolution (sec)	95	98

*average

Tumor segmentation

The T1W images acquired two minutes post-contrast administration were used for the 3D tumor segmentation, as this is generally accepted to be the peak of enhancement of breast cancers [38]. Tumors were independently segmented by four observers with different degrees of experience in breast MR imaging: a dedicated breast radiologist with 11 years of clinical breast MRI experience (ML), a radiology resident with one year of breast MRI clinical experience (TvN), a Ph.D. candidate with a medical degree but no breast MRI clinical experience (RG) and a medical student with no experience whatsoever (NV) (Figure 2). Segmentations were performed manually with Mirada RTx (v1.2.0.59, Mirada Medical, Oxford, UK). Agreements regarding segmentation procedures were made prior to tumor segmentation: (i) observers were allowed to adjust the image grayscale to optimize the visualization of the tumor; (ii) lymph nodes, pectoral muscle, and skin were excluded from segmentation; (iii) spiculae were only segmented if histologically confirmed. All observers had access to the radiology report during segmentation but were blinded to each other's segmentations.

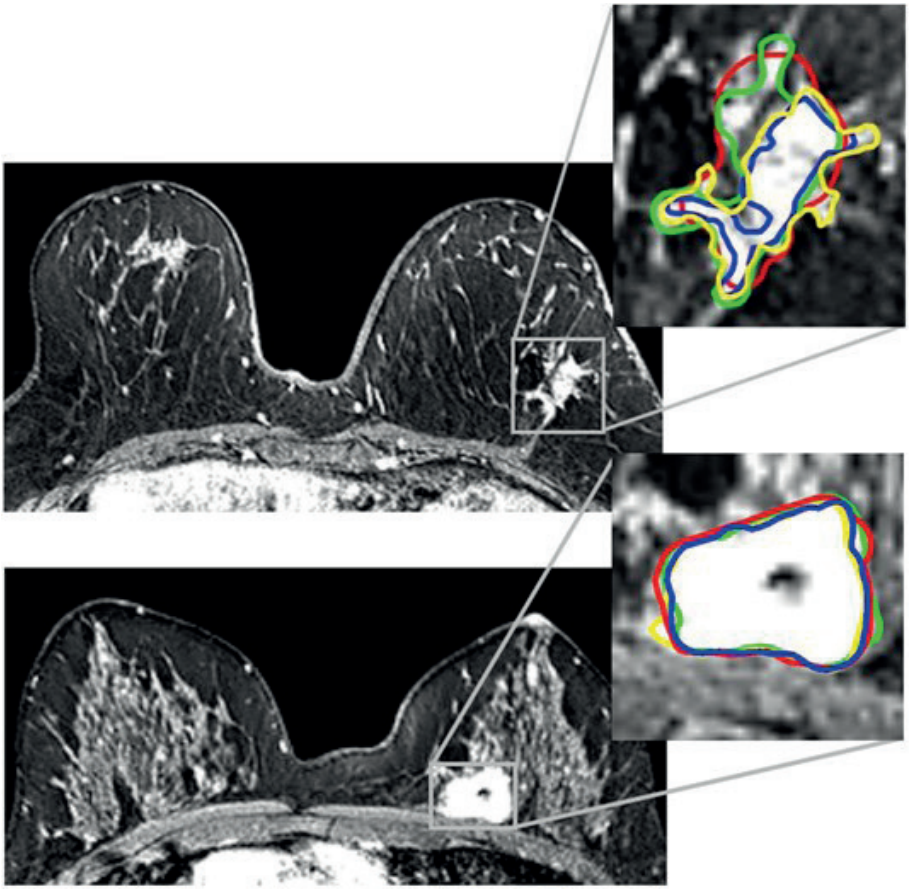


Figure 2. Two invasive breast tumors in the left breast on the 2-min post-contrast DCE-MRI with four single manual segmentations (colored margins: red, blue, green and yellow) fused. Upper: 'challenging tumor' with a mean DSC of 0.78 (range 0.71–0.82). Lower: 'easy tumor' with a mean DSC of 0.90 (range 0.89–0.91).

Image pre-processing and feature extraction

Radiomics feature extraction is generally performed after image pre-processing. Pre-processing is designed to increase data homogeneity, as well as to reduce image noise and computational requirements. Both radiomics software have the optionality to perform image normalization internally before feature extraction, which varies to an extent across the software. Pyradiomics centers the image around the mean and standard deviation based on all gray values of the image, while RadiomiX normalizes the images after removal of background data (non-breast voxels containing air). This transforms the voxel grayscale values to a more comparable range without changing image textures. Each image was discretized by

resampling the grayscale values using a fixed bin width of 0.1 in order to reduce image noise and computational burden. The Pyradiomics community [39] recommends the number of bins to be in range of 16-128. We calculated the optimal bin width by extracting the greyscale ranges within all the ROIs and choosing a width that maximizes the number of ROIs that fall in the abovementioned range of bins. Finally, voxel size was standardized across the cohorts to isotropic 1.0 mm^3 voxels by means of linear interpolation. For each manually segmented ROI, features were extracted using two commonly used radiomics software: RadiomiX Discovery Toolbox software (OncoRadiomics SA, Liège, Belgium) and the open-source Pyradiomics software, version 2.1.2 [39,40]. A mathematical description of all RadiomiX features can be found in supplementary material 5. The Pyradiomics feature description can be found online [41]. Both software is IBSI compliant for most features, with a note being added in case of differences.

Data analysis

Segmentation variability analysis

Features with (near) zero variance across all tumors, i.e. features that have the same value across ninety-five percent or more of the observations, were excluded from the analysis as they carry no discriminative value. To evaluate the variability of the remaining features introduced by manual segmentation, the volumetric Dice Similarity Coefficient (DSC) was calculated for all pairs of observers. The DSC is a metric that quantifies the agreement (or 'overlap') between two segmentations [42]. A DSC of 1 indicates perfect spatial overlap of the segmentations, whereas 0 indicates no agreement, *i.e.* no spatial overlap of the segmentations, and a good overlap is considered with DSC > 0.7 as indicated by the literature [43]. The DSC was calculated as:

$$DSC = 2\frac{(|A \cap B|)}{(|A| + |B|)}$$

where A is the set of voxels contained in the first contour, B is the set of voxels contained in the second contour, || indicates the cardinality of the sets, and ∩ is the intersection between the first and second sets [44]. The DSC was calculated using Python (Version 3.6.3150.1013).

Radiomics feature robustness analysis

Feature robustness was assessed by evaluating the two-way random single measure intraclass correlation coefficient (ICC) (2,1). The two-way random model approach was chosen as it allows generalization of the results to any other rater with similar characteristics [44]. The ICC ranges between 0 and 1, with values closer to 1 representing stronger feature robustness to differences in segmentations. We chose a pre-defined ICC cut-off of >0.9 to select highly stable features that are insensitive to segmentation variability [44]. Feature robustness was calculated for all RadiomiX and Pyradiomics features. The settings for image pre-processing (normalization, discretization, and resampling) in both radiomics software were checked for disparities. Calculations were performed in R studio (version 1.1.456, Vienna, Austria) [45] using the IRR package version 0.84 [46].

Easy- vs. challenging-to-segment tumors analysis

The differences in feature robustness and inter-observer tumor segmentation variability between 'easy-to-segment' and 'challenging-to-segment' tumors ones, hereinafter referred to as 'easy tumors' and 'challenging tumors', were assessed. This classification was unanimously determined by the dedicated breast radiologist (ML). 'Easy tumors' were defined as homogenous, round tumors with relatively sharp (albeit sometimes irregular) margins, without spiculae or areas of accompanying non-mass enhancement. Tumors not meeting these criteria were categorized as 'challenging tumors' (Figure 3). To compare DSC results between 'easy' and 'challenging' tumors we used the independent samples t-test, performed in R studio using the IRR package.

Results

Study population

After the application of inclusion and exclusion criteria, 102 patients were included in the final analysis. Twenty-one of these patients were diagnosed with multifocal breast cancer, bringing the total number of tumors analyzed in this study to 129. Of these, 94 tumors (73%) were assigned 'easy tumors' and the remaining 35 tumors (27%) were assigned 'challenging tumors'. The tumor volume between both groups was significant differently (5.3 vs 10.4 for 'easy and challenging tumors', respectively, p=0.03)

Segmentation variability

DSC distributions of all observer combinations are shown in Figure *3*. The mean DSC was 0.81 (range 0.19-0.96). The mean DSC was higher for the 'easy tumors'

compared to the 'challenging tumors' (0.83 vs. 0.75, respectively, p<0.001). The mean DSC for each observer combination separately, for all tumors, ranged between 0.78 and 0.83, where the segmentations of the breast radiologist and the medical student showed the highest overlap.

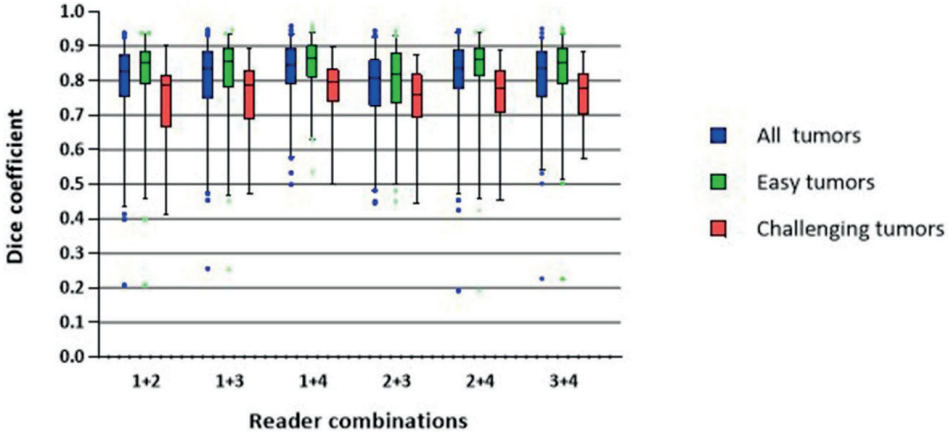


Figure 3. Tumor segmentation variability for pairwise comparison of the different observers. (1) Dedicated breast radiologist, (2) Radiology resident, (3) Ph.D. candidate with a medical degree and (4) Medical student.

Pre-processing and feature extraction

The bin width for image discretization (calculated from the ROI greyscale range) was 0.1. Discretization of the scans with bins 0.1 wide resulted in a mean of 61 grayscale values per image (range 27 -131). RadiomiX and Pyradiomics software extracted a total of 1328 and 833 features for each ROI, respectively. The extracted radiomics features included shape features, first-order statistical, intensity-histogram based, fractal, local intensity, and texture matrix-based features from both unfiltered and filtered images (wavelet decompositions). The RadiomiX software extracts more feature groups compared to the Pyradiomics software, namely intensity histogram (IH), fractal, local intensity, and gray level dependency zone matrix (GLDZM) features.

Radiomics feature robustness

The average ICC for all RadiomiX features was 0.86 (95% CI: 0.85-0.86) and for all Pyradiomics features 0.84 (95% CI: 0.83-0.84). Table 2 presents the average ICC value per feature group for both software. The local intensity features scored the highest average ICC value for the RadiomiX features, and the first-order statistical features score the highest average ICC for the Pyradiomics features.

Table 2. Average ICC values per feature group of the unfiltered and wavelet RadiomiX and Pyradiomics features

	OncoRadiomiX		Pyradiomics	
Feature group (n)	**Mean ICC**	**Range**	**Mean ICC**	**Range**
Shape	0.79	0.57 – 0.93	0.80	0.69 – 0.92
Signal intensity First-order statistics	0.85	0.51 – 0.99	0.84	0.50 – 0.97
IH	0.76	0.63 – 0.98	-	-
Fractal	0.81	0.79 – 0.83	-	-
LocInt	0.95	0.93 – 0.96	-	-
GLCM	0.76	0.49 – 0.88	0.80	0.71 – 0.88
GLRLM	0.79	0.56 – 0.96	0.81	0.63 – 0.95
GLSZM	0.80	0.55 – 0.98	0.84	0.58 – 0.97
GLDZM	0.76	0.50 – 0.92	-	-
NGTDM	0.78	0.57 – 0.85	0.80	0.72 – 0.91
(N)GLDM	0.83	0.55 – 0.96	0.79	0.52 – 0.96
Wavelet	0.81	0.01 – 0.99	0.81	0.12 – 0.99

The percentage of features that scored an ICC > 0.90, and thus were labeled by our pre-determined ICC cut-off as robust, was 41.6% (552/1328) for RadiomiX features and 32.8% (273/833) for Pyradiomics features. The unfiltered RadiomiX features (*i.e.*, calculated on the unfiltered images) had an average ICC value of 0.79 (95% CI: 0.77 – 0.81), of which 41.1% (69/168) were robust (Figure 4). The unfiltered Pyradiomics features had an average ICC value of 0.81 (95% CI: 0.79-0.83), of which 16.2% (17/105) were robust (Figure 5). The results of the wavelet feature groups for both software are presented in the supplementary material 1 and 2.

The percentage of robust RadiomiX features for the 'easy tumors' and the 'challenging tumors' was 57.5% (763/1328) and 17.2% (228/1328), respectively. When only considering the 168 unfiltered features, 50.0% (84/168) of the 'easy tumors' were robust and 20.2% (34/168) of the 'challenging tumors' (supplementary material 3). The percentage of robust Pyradiomics features for the 'easy tumors' and the 'challenging tumors' was 35.7% (297/833) and 28.6% (238/833), respectively. When only considering the 105 unfiltered features, 23.8% (25/105) of the 'easy tumors' were robust and 14.3% (15/105) of the 'challenging tumors' (supplementary material 4).

Figure 4. ICC values of all unfiltered RadiomiX features with robust features (ICC > 0.90) shown in green.

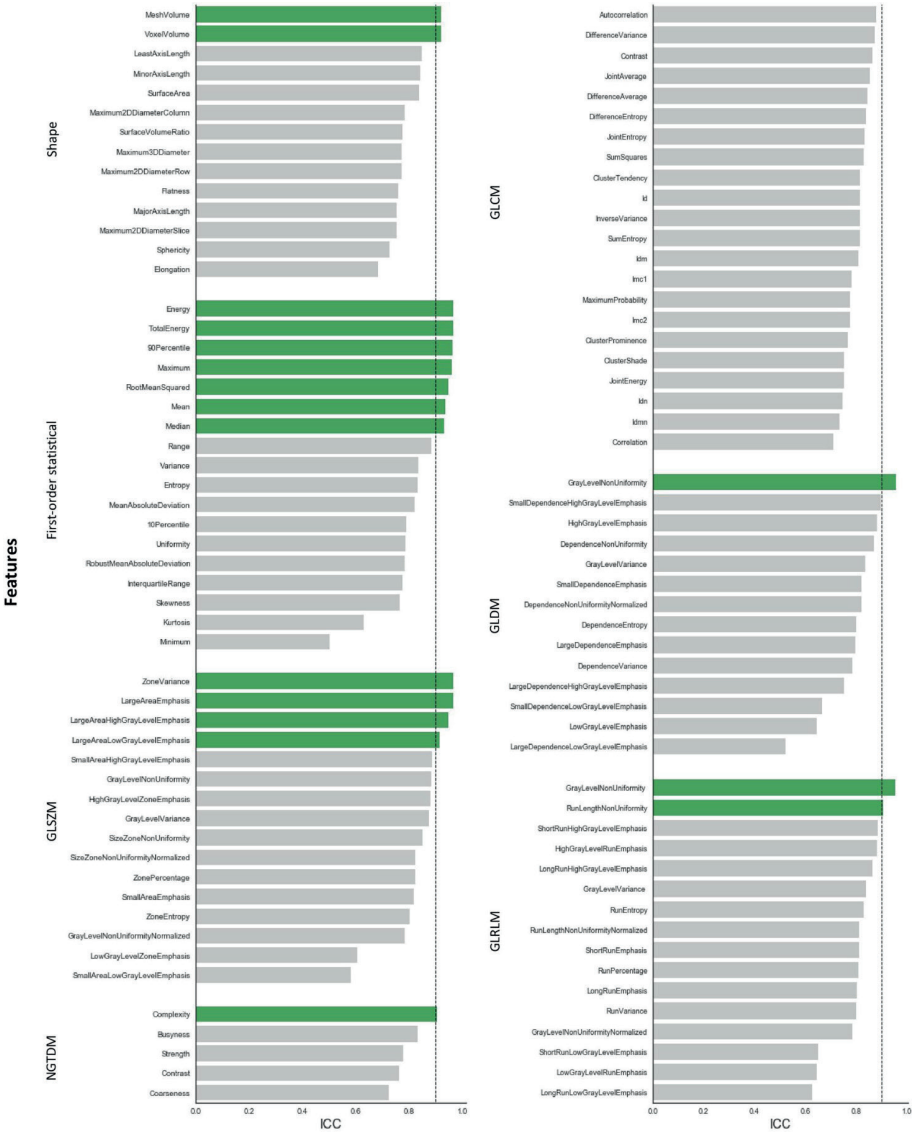


Figure 5. ICC values of all unfiltered Pyradiomics features with robust features (ICC > 0.90) shown in green.

Discussion

In this study, our ultimate goal was to define a list of robust MRI radiomics features, independent of inter-observer segmentation variability, which could facilitate further breast MRI-based radiomics research. We successfully identified a subgroup of robust features for two commonly used radiomics software (41.6% of all RadiomiX features and 32.8% of all Pyradiomics features) in the presence of inter-observer segmentation variability (mean DSC of 0.81).

Although MRI feature robustness has already been investigated for different tumor sites (*e.g.,* cervical cancer [19] and glioblastoma [23]), the effect of inter-observer variability segmentation is most likely tumor-site specific [47]. The feature groups enclosing the most robust features in previous investigations (shape [19] and, Intensity-histogram and GLCM [23]) are different from what we found to be the feature group enclosing the most robust features (local intensities and GLRLM). Most likely this could be explained that different tumor sites influence inter-observer variability. Although one must not forget that the differences in MRI sequences and, feature extraction software also influence this variability. Therefore, the MRI feature robustness cannot be generalized and must be examined for each specific tumor site, taking into account different MRI sequences and feature extraction software.

In addition, feature robustness for both radiomics software was identified for 'easy tumors' and 'challenging tumors'. The number of robust features increased for 'easy tumors' and decreased for 'challenging tumors' in both software with significant differences between the mean DSC of the 'easy' and 'challenging' tumors (0.83 vs. 0.75, respectively, $p < 0.001$). The fact that the 'challenging tumors' were more irregular, often with spiculae, causes more segmentation variability and therefore less robust features. Furthermore, the significant difference in the DSC between easy and challenging tumors could be attributed to the sensitivity of the metric to tumor volume. Easy tumors were on average significantly smaller than challenging ones; therefore, a minor difference in segmentation of a small tumor would have a more profound effect on the DSC, compared to those with larger volumes.

A detailed comparison to previous studies is limited to one similar study. Saha et al [48] investigated the impact of breast MRI segmentation variability on radiomics feature robustness, whereby features were extracted using in-house software. Their reported mean ICC of 0.85 for all features, using semi-automatic breast tumor segmentation, is comparable to the average ICC reported in this study. Although the segmentations were performed by four fellow breast radiology trainees, the DSC results they report (range: 0.506-0.740) were much lower than the DSC results in our analysis (range: 0.783-0.827). We consciously opted for people with different segmentation expertise to ensure observer-independence

of the robust features, consequently widening the applicability. Approximately 10% of the tumor features in their article were found to be robust, compared to 41.1% in this study. Solely 20 textural features (GLCM) were comparable between the studies, whereby the ICC of these features showed a substantial difference (average 0.26, range 0.09 – 0.51).

While we present the robust features for two different radiomics software, our aim is solely to facilitate future application of our findings. Both software have different pre-processing steps, and different groups of features, and comparing the software is beyond the scope of this study. A global initiative to standardize radiomic features extraction using different radiomics software–Imaging Biomarkers Standardization Initiative (IBSI)- was started to address these issues in a more comprehensive fashion [49].

To overcome the problem of inter-observer variability with respect to ROI segmentation, promising steps towards (semi-)automatic segmentation have been taken in other tumor sites [50-54]. However, little work has been published on fully automatic segmentation software for DCE-MRI of the breast [33,55-57]. Most software, including semi-automatic segmentation, still require manual input or adjustments [33,55,56], and would still be significantly slower than fully automated segmentation. Recent work on automatic MRI breast tissue segmentation reported encouraging results but was performed on only 30 patients [57]. The current lack of reliable, validated and widely available automatic segmentation software tools, and the need for manual input in semi-automated segmentation, demonstrate that manual segmentation remains important. The use of protocols or guidelines could encourage more reproducible manual segmentation results [58,59]. Furthermore, by providing precise instructions before the start of segmentation, inter-observer segmentation variability can be minimized.

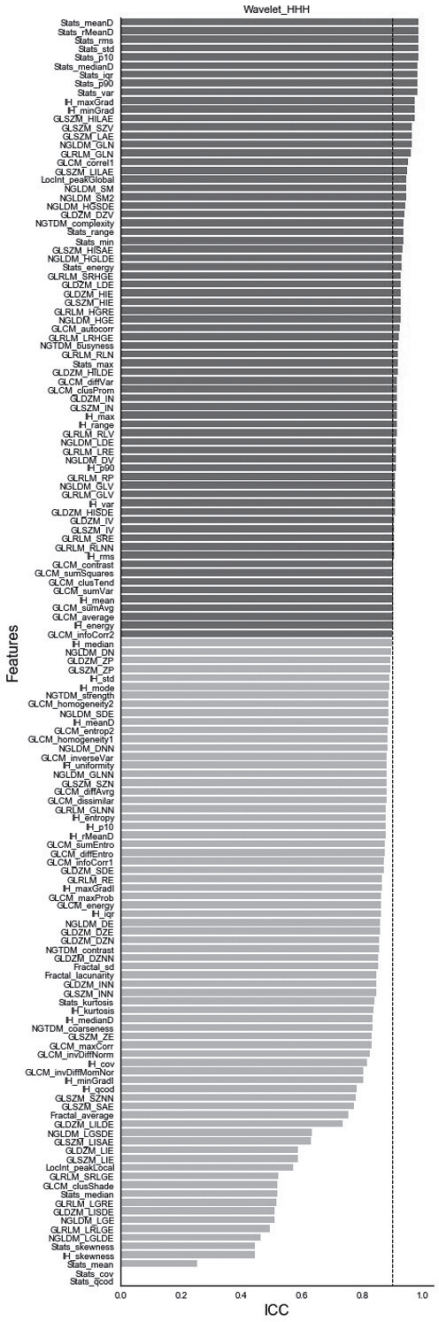
There are some limitations to this study. Although an ICC threshold value of 0.90 was chosen to determine feature robustness, the significance of this threshold for radiomics models for patients' outcome prediction is yet to be investigated. The inclusion of more patients and observers will allow better generalization of the results and development of robust radiomics signatures. Furthermore, we identified feature robustness to segmentation observer variability. However, due to the lack of data, we were not able to assess the robustness of radiomics features to differences in image acquisition, pre-processing and feature extraction, which are other major challenges in radiomics analysis. These are the aim of our current studies.

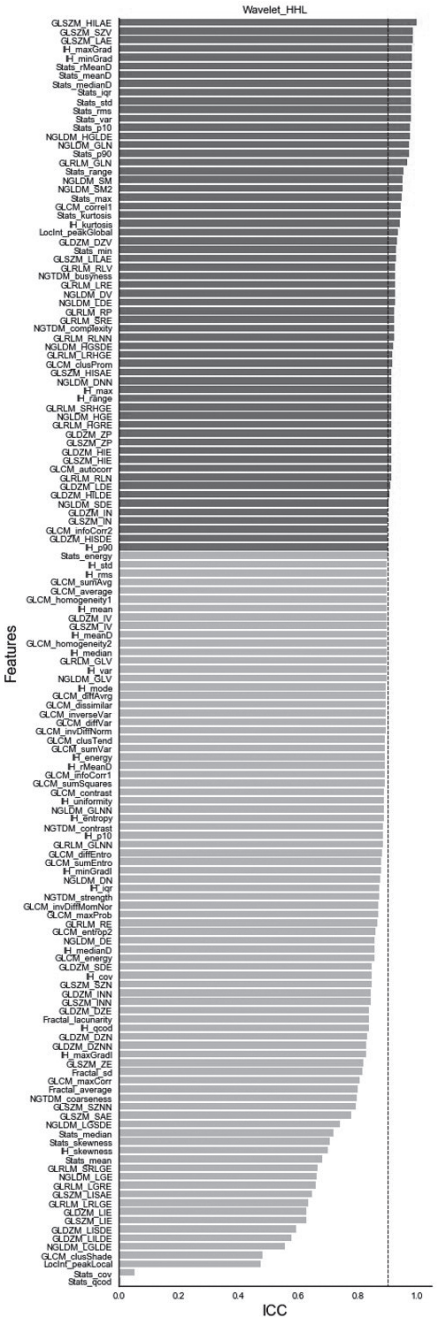
In conclusion, this study shows the intuitive notion that more complex, challenging tumors lead to less robust features. We identified radiomics features robust to inter-observer variations across two different radiomics software, which could

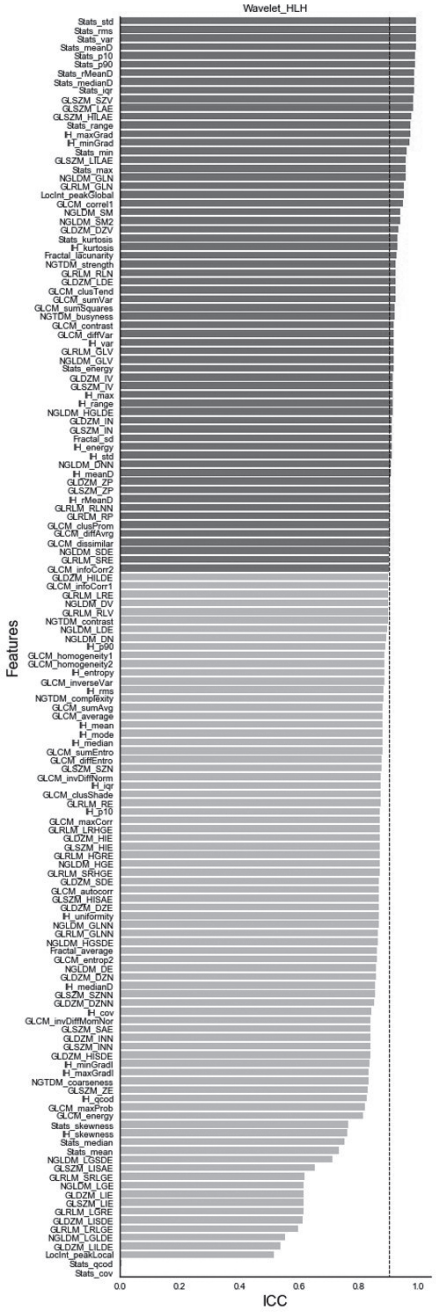
be used for preselection of radiomics features in future radiomics analysis concerning MRI-based breast radiomics. Ultimately, this study identified a list of robust radiomics features, which is independent of inter-observer segmentation variability in breast MRI for two commonly used software.

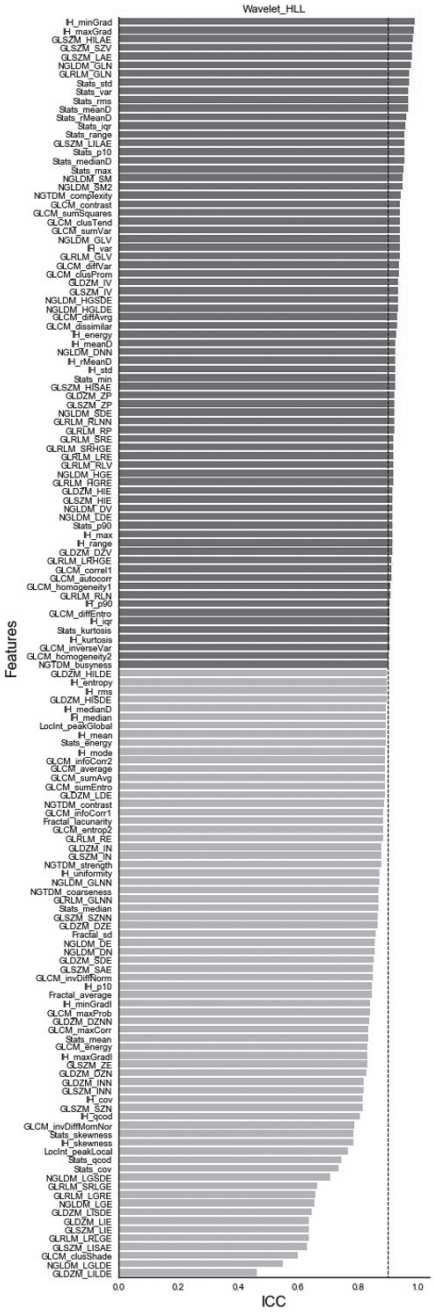
Supplementary materials

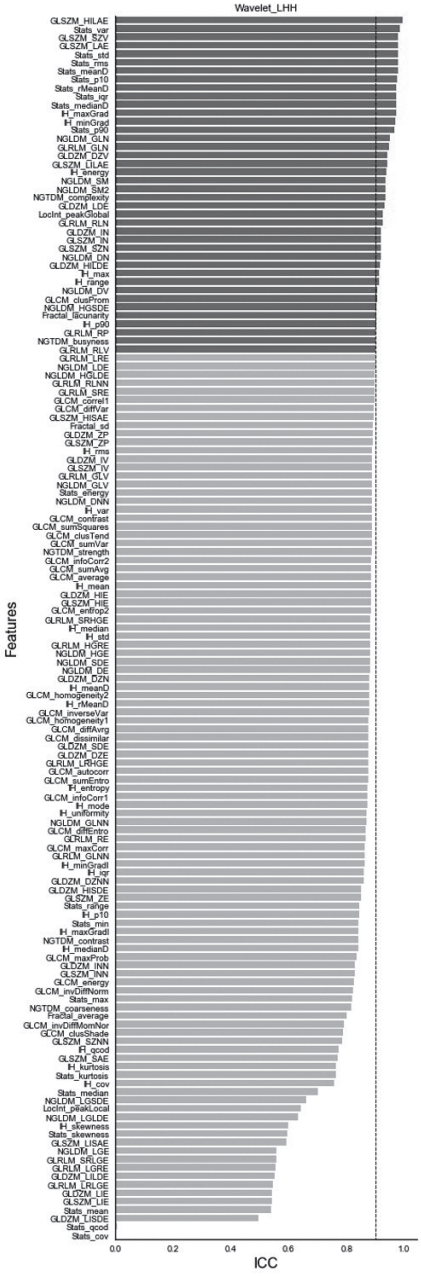
Supplementary material 1. ICC values of all wavelet RadiomiX features per wavelet decomposition with robust features (ICC >0.9) shown in dark gray

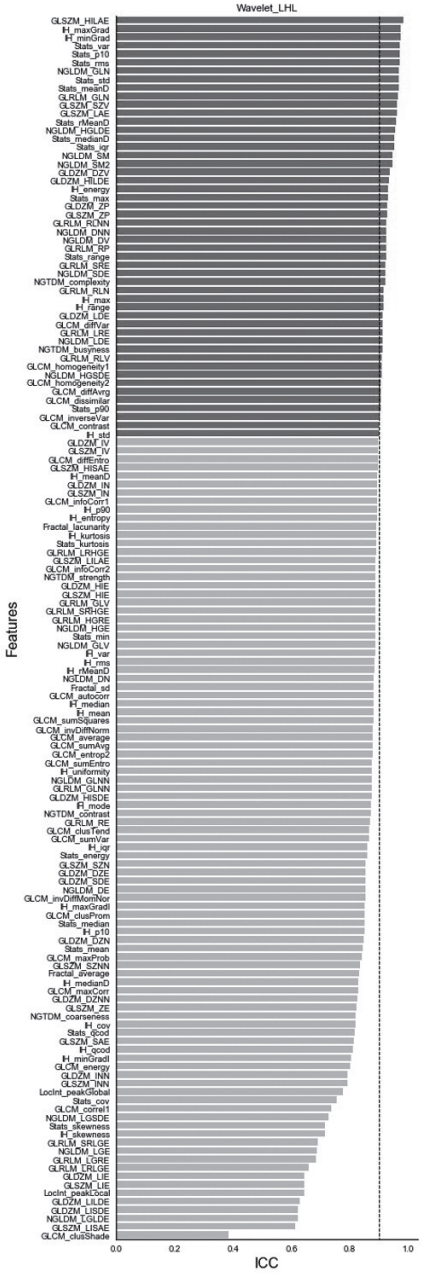


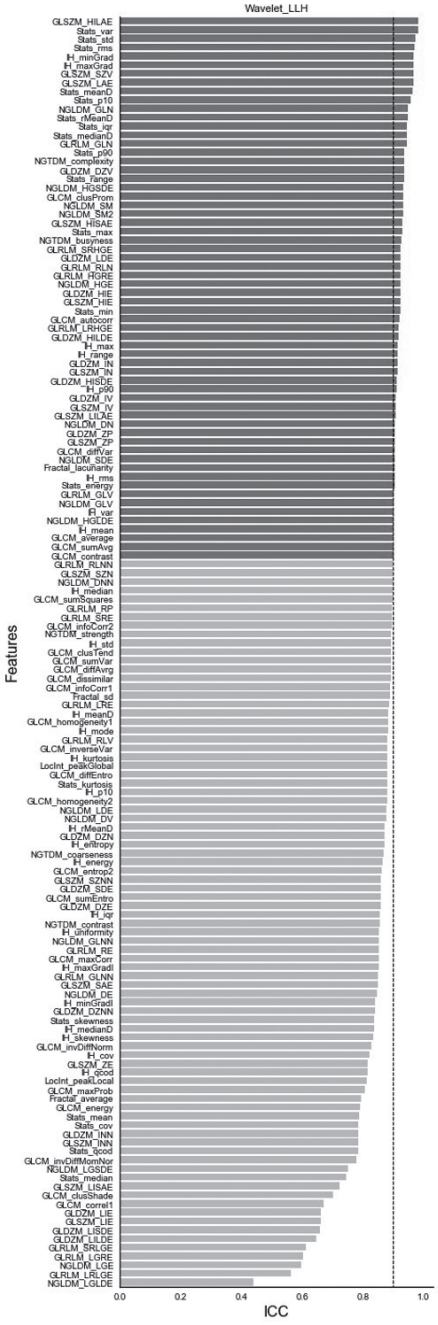


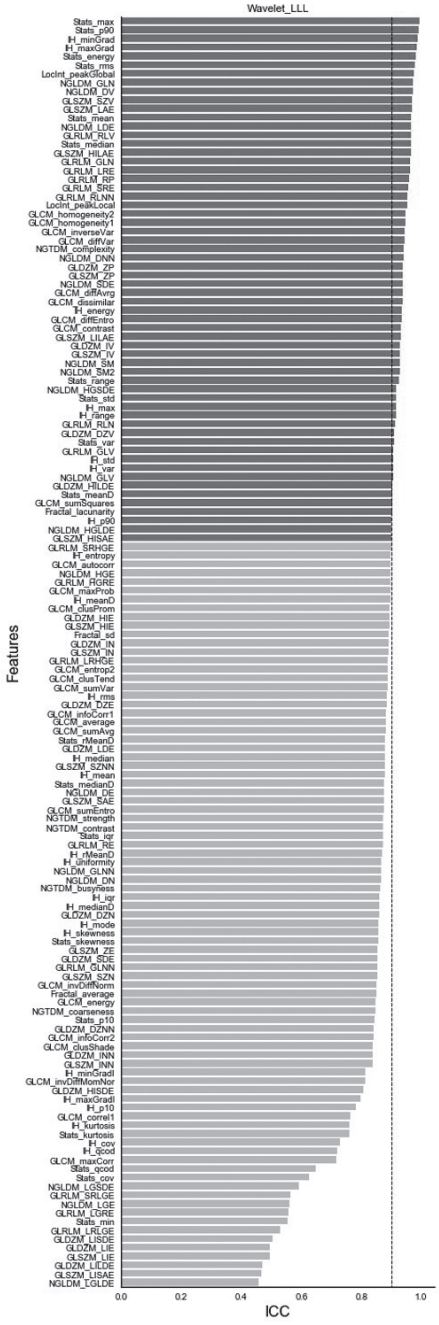

Wavelet_HLH
Features
ICC
0.0
0.2
0.4
0.6
0.8
1.0


Wavelet_HLL
Features
ICC
0.0
0.2
0.4
0.6
0.8
1.0

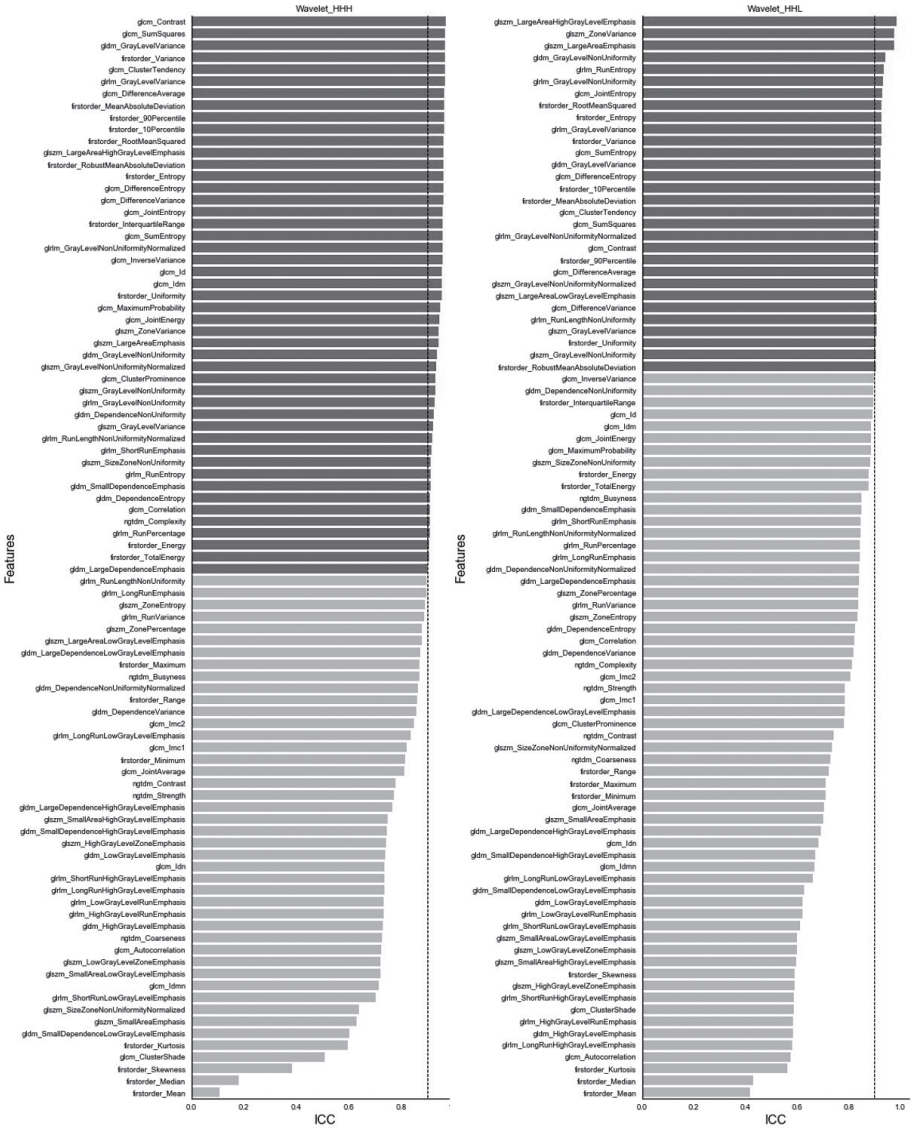

Wavelet_LHH
Features
GLSZM_HILAE
Stats_var
GLSZM_SZV
GLSZM_LAE
Stats_std
Stats_rms
Stats_meanD
Stats_p10
Stats_rMeanD
Stats_iqr
Stats_medianD
IH_maxGrad
IH_minGrad
Stats_p90
NGLDM_GLN
GLRLM_GLN
GLDZM_DZV
GLSZM_LILAE
IH_energy
NGLDM_SM
NGLDM_SM2
NGTDM_complexity
GLDZM_LDE
LocInt_peakGlobal
GLRLM_RLN
GLDZM_IN
GLSZM_IN
GLSZM_SZN
NGLDM_DN
GLDZM_HILDE
IH_max
IH_range
NGLDM_DV
GLCM_clusProm
NGLDM_HGSDE
Fractal_lacunarity
IH_p90
GLRLM_RP
NGTDM_busyness
GLRLM_RLV
GLRLM_LRE
NGLDM_LDE
NGLDM_HGLDE
GLRLM_RLNN
GLRLM_SRE
GLCM_correl1
GLCM_diffVar
GLSZM_HISAE
Fractal_sd
GLDZM_ZP
GLSZM_ZP
IH_rms
GLDZM_IV
GLSZM_IV
GLRLM_GLV
NGLDM_GLV
Stats_allenrgy
NGLDM_DNN
IH_var
GLCM_contrast
GLCM_sumSquares
GLCM_clusTend
GLCM_sumVar
NGTDM_strength
GLCM_infoCorr2
GLCM_sumAvg
GLCM_average
IH_mean
GLDZM_HIE
GLSZM_HIE
GLCM_entrop2
GLRLM_SRHGE
IH_median
IH_std
GLRLM_HGRE
NGLDM_HGE
NGLDM_SDE
NGLDM_DE
GLDZM_DZN
IH_meanD
GLCM_homogeneity2
IH_rMeanD
GLCM_inverseVar
GLCM_homogeneity1
GLCM_diffAvrg
GLCM_dissimilar
GLDZM_SDE
GLDZM_DZE
GLRLM_LRHGE
GLCM_autocorr
GLCM_sumEntro
IH_entropy
GLCM_infoCorr1
IH_mode
IH_uniformity
NGLDM_GLNN
GLCM_diffEntro
GLRLM_RE
GLCM_maxCorr
GLRLM_GLNN
IH_minGradI
IH_iqr
GLDZM_DZNN
GLDZM_HISDE
GLSZM_ZE
Stats_range
IH_p10
Stats_min
IH_maxGradI
NGTDM_contrast
IH_medianD
GLCM_maxProb
GLDZM_INN
GLSZM_INN
GLCM_energy
GLCM_invDiffMomNor
Stats_max
NGTDM_coarseness
Fractal_average
GLCM_invDiffMomNor
GLCM_clusShade
GLSZM_SZNN
IH_qcod
GLSZM_SAE
IH_kurtosis
Stats_kurtosis
IH_cov
Stats_median
NGLDM_LGSDE
LocInt_peakLocal
NGLDM_LGLDE
IH_skewness
Stats_skewness
GLSZM_LISAE
NGLDM_LGE
GLRLM_SRLGE
GLRLM_LGRE
GLDZM_LILDE
GLRLM_LRLGE
GLDZM_LIE
GLSZM_LIE
Stats_mean
GLDZM_LISDE
Stats_qcod
Stats_cov
0.0
0.2
0.4
0.6
0.8
1.0
ICC

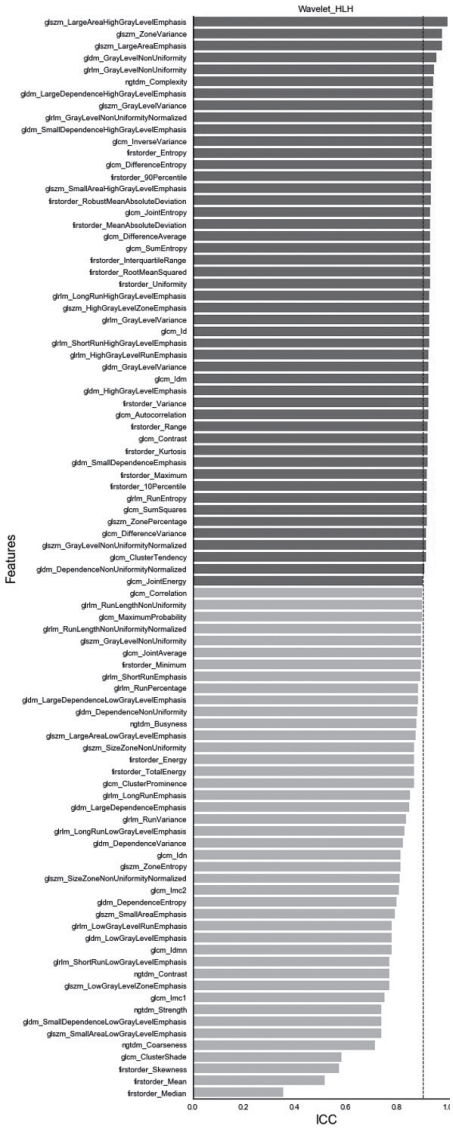

Wavelet_LHL
Features
GLSZM_HILAE
IH_maxGrad
IH_minGrad
Stats_var
Stats_p10
Stats_rms
NGLDM_GLN
Stats_std
Stats_meanD
GLRLM_GLN
GLSZM_SZV
GLSZM_LAE
Stats_rMeanD
NGLDM_HGLDE
Stats_medianD
Stats_iqr
NGLDM_SM
NGLDM_SM2
GLDZM_DZV
GLDZM_HILDE
IH_energy
Stats_max
GLDZM_ZP
GLSZM_ZP
GLRLM_RLNN
NGLDM_DNN
NGLDM_DV
GLRLM_RP
Stats_range
GLRLM_SRE
NGLDM_SDE
NGTDM_complexity
GLRLM_RLN
IH_max
IH_range
GLDZM_LDE
GLCM_diffVar
GLRLM_LRE
NGLDM_LDE
NGTDM_busyness
GLRLM_RLV
GLCM_homogeneity1
NGLDM_HGSDE
GLCM_homogeneity2
GLCM_diffAvrg
GLCM_dissimilar
Stats_p90
GLCM_inverseVar
GLCM_contrast
IH_std
GLDZM_IV
GLSZM_IV
GLCM_diffEntro
GLSZM_HISAE
IH_meanD
GLDZM_IN
GLSZM_IN
GLCM_infoCorr1
IH_p90
IH_entropy
Fractal_lacunarity
IH_kurtosis
Stats_kurtosis
GLRLM_LRHGE
GLSZM_LILAE
GLCM_infoCorr2
NGTDM_strength
GLDZM_HIE
GLSZM_HIE
GLRLM_GLV
GLRLM_SRHGE
GLRLM_HGRE
NGLDM_HGE
Stats_min
NGLDM_GLV
IH_var
IH_rms
IH_rMeanD
NGLDM_DN
Fractal_sd
GLCM_autocorr
IH_median
IH_mean
GLCM_sumSquares
GLCM_invDiffNorm
GLCM_average
GLCM_sumAvg
GLCM_entrop2
GLCM_sumEntro
IH_uniformity
NGLDM_GLNN
GLRLM_GLNN
GLDZM_HISDE
IH_mode
NGTDM_contrast
GLRLM_RE
GLCM_clusTend
GLCM_sumVar
IH_iqr
Stats_energy
GLSZM_SZN
GLDZM_DZE
GLDZM_SDE
NGLDM_DE
GLCM_invDiffMomNor
IH_maxGradI
GLCM_clusProm
Stats_median
IH_p10
GLDZM_DZN
Stats_mean
GLCM_maxProb
GLSZM_SZNN
Fractal_average
IH_medianD
GLCM_maxCorr
GLDZM_DZNN
GLSZM_ZE
NGTDM_coarseness
IH_cov
Stats_qcod
GLSZM_SAE
IH_qcod
IH_minGradI
GLCM_energy
GLDZM_INN
GLSZM_INN
LocInt_peakGlobal
Stats_cov
GLCM_correl1
NGLDM_LGSDE
Stats_skewness
IH_skewness
GLRLM_SRLGE
NGLDM_LGE
GLRLM_LGRE
GLRLM_LRLGE
GLDZM_LIE
GLSZM_LIE
LocInt_peakLocal
GLDZM_LILDE
GLDZM_LISDE
NGLDM_LGLDE
GLSZM_LISAE
GLCM_clusShade
0.0
0.2
0.4
0.6
0.8
1.0
ICC

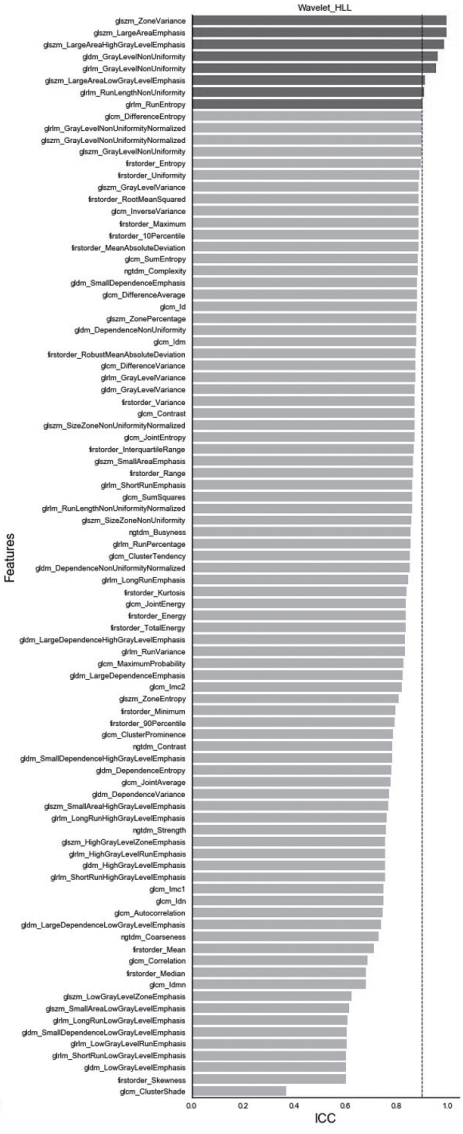

Wavelet_LLH
Features
ICC
0.0
0.2
0.4
0.6
0.8
1.0
GLSZM_HILAE
Stats_var
Stats_std
Stats_rms
IH_minGrad
IH_maxGrad
GLSZM_SZV
GLSZM_LAE
Stats_meanD
Stats_p10
NGLDM_GLN
Stats_rMeanD
Stats_iqr
Stats_medianD
GLRLM_GLN
Stats_p90
NGTDM_complexity
GLDZM_DZV
Stats_range
NGLDM_HGSDE
GLCM_clusProm
NGLDM_SM
NGLDM_SM2
GLSZM_HISAE
Stats_max
NGTDM_busyness
GLRLM_SRHGE
GLDZM_LDE
GLRLM_RLN
GLRLM_HGRE
NGLDM_HGE
GLDZM_HIE
GLSZM_HIE
Stats_min
GLCM_autocorr
GLRLM_LRHGE
GLDZM_HILDE
IH_max
IH_range
GLDZM_IN
GLSZM_IN
GLDZM_HISDE
IH_p90
GLDZM_IV
GLSZM_IV
GLSZM_LILAE
NGLDM_DN
GLDZM_ZP
GLSZM_ZP
GLCM_diffVar
NGLDM_SDE
Fractal_lacunarity
IH_rms
Stats_energy
GLRLM_GLV
NGLDM_GLV
IH_var
NGLDM_HGLDE
IH_mean
GLCM_average
GLCM_sumAvg
GLCM_contrast
GLRLM_RLNN
GLSZM_SZN
NGLDM_DNN
IH_median
GLCM_sumSquares
GLRLM_RP
GLRLM_SRE
GLCM_infoCorr2
NGTDM_strength
IH_std
GLCM_clusTend
GLCM_sumVar
GLCM_diffAvg
GLCM_dissimilar
GLCM_infoCorr1
Fractal_sd
GLRLM_LRE
IH_meanD
GLCM_homogeneity1
IH_mode
GLRLM_RLV
GLCM_inverseVar
IH_kurtosis
LocInt_peakGlobal
GLCM_diffEntro
Stats_kurtosis
IH_p10
GLCM_homogeneity2
NGLDM_LDE
NGLDM_DV
IH_rMeanD
GLDZM_DZN
IH_entropy
NGTDM_coarseness
IH_energy
GLCM_entro2
GLSZM_SZNN
GLDZM_SDE
GLCM_sumEntro
GLDZM_DZE
IH_iqr
NGTDM_contrast
IH_uniformity
NGLDM_GLNN
GLRLM_RE
GLCM_maxCorr
IH_maxGradl
GLRLM_GLNN
GLSZM_SAE
NGLDM_DE
IH_minGradl
GLDZM_DZNN
Stats_skewness
IH_medianD
IH_skewness
GLCM_invDiffNorm
IH_cov
GLSZM_ZE
IH_qcod
LocInt_peakLocal
GLCM_maxProb
Fractal_average
GLCM_energy
Stats_mean
Stats_cov
GLDZM_INN
GLSZM_INN
Stats_qcod
GLCM_invDiffMomNor
NGLDM_LGSDE
Stats_median
GLSZM_LISAE
GLCM_clusShade
GLCM_correl1
GLDZM_LIE
GLSZM_LIE
GLDZM_LISDE
GLDZM_LILDE
GLRLM_SRLGE
GLRLM_LGRE
NGLDM_LGE
GLRLM_LRLGE
NGLDM_LGLDE


Wavelet_LLL
Features
ICC
0.0
0.2
0.4
0.6
0.8
1.0
Stats_max
Stats_p90
IH_minGrad
IH_maxGrad
Stats_energy
Stats_rms
LocInt_peakGlobal
NGLDM_GLN
NGLDM_DV
GLSZM_SZV
GLSZM_LAE
Stats_mean
NGLDM_LDE
GLRLM_RLV
Stats_median
GLSZM_HILAE
GLRLM_GLN
GLRLM_LRE
GLRLM_RP
GLRLM_SRE
GLRLM_RLNN
LocInt_peakLocal
GLCM_homogeneity2
GLCM_homogeneity1
GLCM_inverseVar
GLCM_diffVar
NGTDM_complexity
NGLDM_DNN
GLDZM_ZP
GLSZM_ZP
NGLDM_SDE
GLCM_diffAvg
GLCM_dissimilar
IH_energy
GLCM_diffEntro
GLCM_contrast
GLSZM_LILAE
GLDZM_IV
GLSZM_IV
NGLDM_SM
NGLDM_SM2
Stats_range
NGLDM_HGSDE
Stats_std
IH_max
IH_range
GLRLM_RLN
GLDZM_DZV
Stats_var
GLRLM_GLV
IH_std
IH_var
NGLDM_GLV
GLDZM_HILDE
Stats_meanD
GLCM_sumSquares
Fractal_lacunarity
IH_p90
NGLDM_HGLDE
GLSZM_HISAE
GLRLM_SRHGE
IH_entropy
GLCM_autocorr
NGLDM_HGE
GLRLM_HGRE
GLCM_maxProb
IH_meanD
GLCM_clusProm
GLDZM_HIE
GLSZM_HIE
Fractal_sd
GLDZM_IN
GLSZM_IN
GLRLM_LRHGE
GLCM_entrop2
GLCM_clusTend
GLCM_sumVar
IH_rms
GLDZM_DZE
GLCM_infoCorr1
GLCM_average
GLCM_sumAvg
Stats_rMeanD
GLDZM_LDE
IH_median
GLSZM_SZNN
IH_mean
Stats_medianD
NGLDM_DE
GLSZM_SAE
GLCM_sumEntro
NGTDM_strength
NGTDM_contrast
Stats_iqr
GLRLM_RE
IH_rMeanD
IH_uniformity
NGLDM_GLNN
NGLDM_DN
NGTDM_busyness
IH_iqr
IH_medianD
GLDZM_DZN
IH_mode
IH_skewness
Stats_skewness
GLSZM_ZE
GLDZM_SDE
GLRLM_GLNN
GLSZM_SZN
GLCM_invDiffNorm
Fractal_average
GLCM_energy
NGTDM_coarseness
Stats_p10
GLDZM_DZNN
GLCM_infoCorr2
GLCM_clusShade
GLDZM_INN
GLSZM_INN
IH_minGradl
GLCM_invDiffMomNor
GLDZM_HISDE
IH_maxGradl
IH_p10
GLCM_correl1
IH_kurtosis
Stats_kurtosis
IH_cov
IH_qcod
GLCM_maxCorr
Stats_qcod
Stats_cov
NGLDM_LGSDE
GLRLM_SRLGE
NGLDM_LGE
GLRLM_LGRE
Stats_min
GLRLM_LRLGE
GLDZM_LISDE
GLDZM_LIE
GLSZM_LIE
GLDZM_LILDE
GLSZM_LISAE
NGLDM_LGLDE

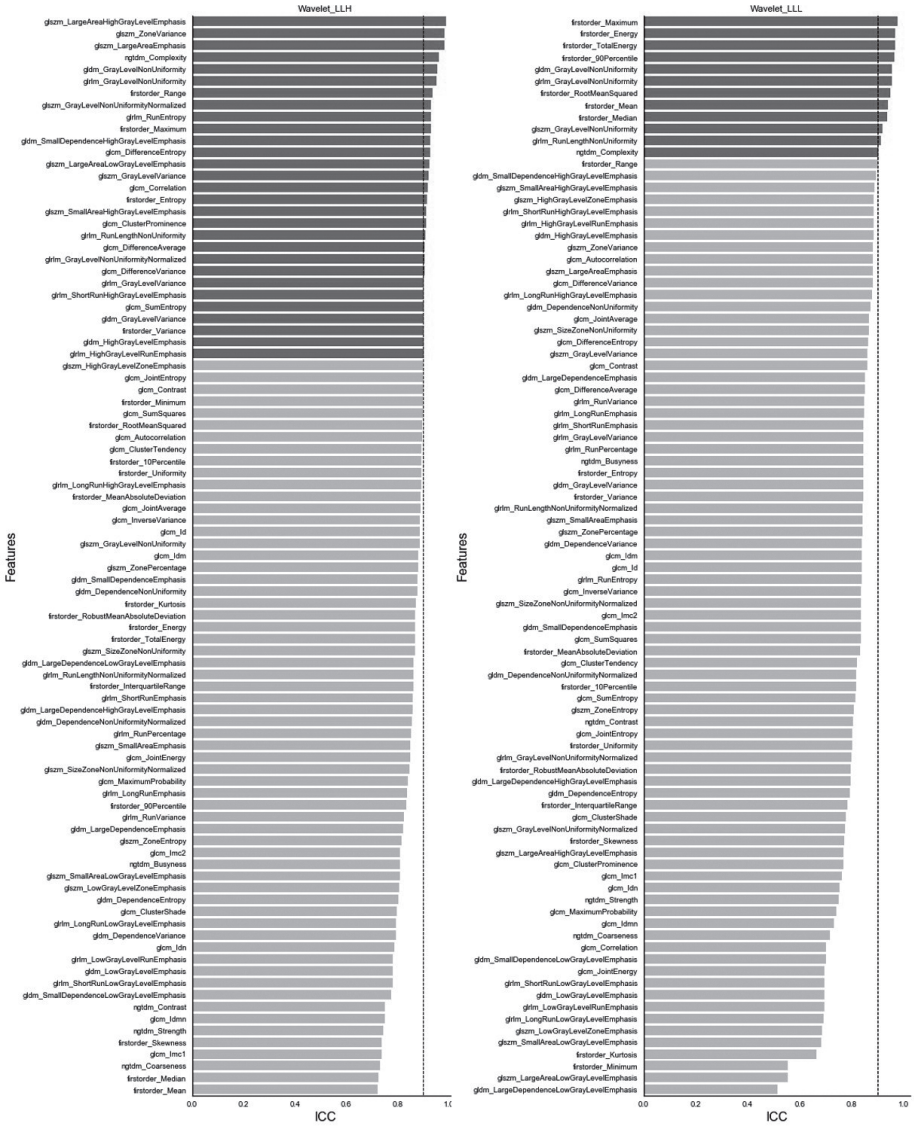
Supplementary material 2. ICC values of all wavelet Pyradiomics features per wavelet decomposition with robust features (ICC >0.9) shown in dark gray



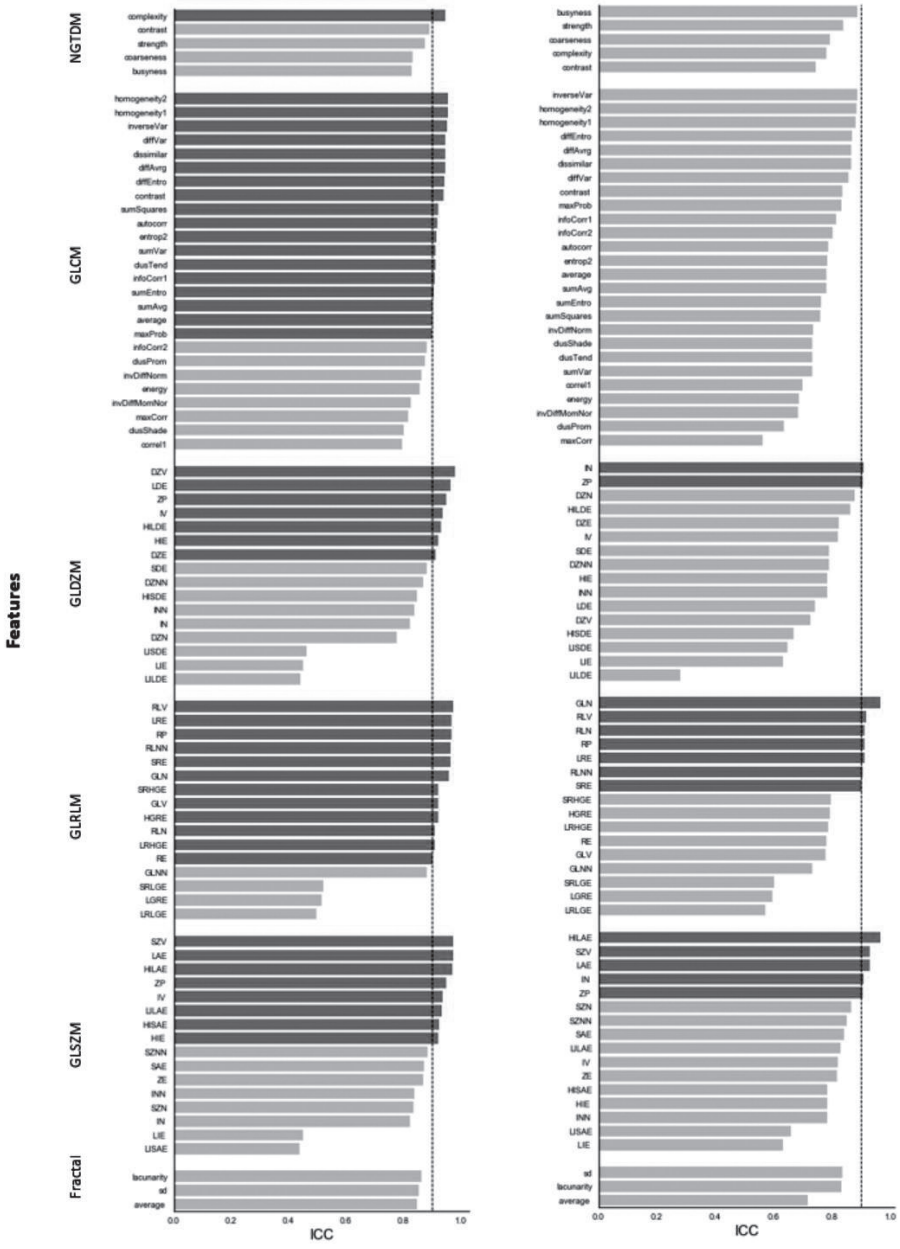

Wavelet_HLH
glszm_LargeAreaHighGrayLevelEmphasis
glszm_ZoneVariance
glszm_LargeAreaEmphasis
gldm_GrayLevelNonUniformity
glrlm_GrayLevelNonUniformity
ngtdm_Complexity
gldm_LargeDependenceHighGrayLevelEmphasis
glszm_GrayLevelVariance
glrlm_GrayLevelNonUniformityNormalized
gldm_SmallDependenceHighGrayLevelEmphasis
glcm_InverseVariance
firstorder_Entropy
glcm_DifferenceEntropy
firstorder_90Percentile
glszm_SmallAreaHighGrayLevelEmphasis
firstorder_RobustMeanAbsoluteDeviation
glcm_JointEntropy
firstorder_MeanAbsoluteDeviation
glcm_DifferenceAverage
glcm_SumEntropy
firstorder_InterquartileRange
firstorder_RootMeanSquared
firstorder_Uniformity
glrlm_LongRunHighGrayLevelEmphasis
glszm_HighGrayLevelZoneEmphasis
glrlm_GrayLevelVariance
glcm_Id
glrlm_ShortRunHighGrayLevelEmphasis
glrlm_HighGrayLevelRunEmphasis
gldm_GrayLevelVariance
glcm_Idm
gldm_HighGrayLevelEmphasis
firstorder_Variance
glcm_Autocorrelation
firstorder_Range
glcm_Contrast
firstorder_Kurtosis
gldm_SmallDependenceEmphasis
firstorder_Maximum
firstorder_10Percentile
glrlm_RunEntropy
glcm_SumSquares
glszm_ZonePercentage
glcm_DifferenceVariance
glszm_GrayLevelNonUniformityNormalized
glcm_ClusterTendency
gldm_DependenceNonUniformityNormalized
glcm_JointEnergy
glcm_Correlation
glrlm_RunLengthNonUniformity
glcm_MaximumProbability
glrlm_RunLengthNonUniformityNormalized
glszm_GrayLevelNonUniformity
glcm_JointAverage
firstorder_Minimum
glrlm_ShortRunEmphasis
glrlm_RunPercentage
gldm_LargeDependenceLowGrayLevelEmphasis
gldm_DependenceNonUniformity
ngtdm_Busyness
glszm_LargeAreaLowGrayLevelEmphasis
glszm_SizeZoneNonUniformity
firstorder_Energy
firstorder_TotalEnergy
glcm_ClusterProminence
glrlm_LongRunEmphasis
gldm_LargeDependenceEmphasis
glrlm_RunVariance
glrlm_LongRunLowGrayLevelEmphasis
gldm_DependenceVariance
glcm_Idn
glszm_ZoneEntropy
glszm_SizeZoneNonUniformityNormalized
glcm_Imc2
gldm_DependenceEntropy
glszm_SmallAreaEmphasis
glrlm_LowGrayLevelRunEmphasis
gldm_LowGrayLevelEmphasis
glcm_Idmn
glrlm_ShortRunLowGrayLevelEmphasis
ngtdm_Contrast
glszm_LowGrayLevelZoneEmphasis
glcm_Imc1
ngtdm_Strength
gldm_SmallDependenceLowGrayLevelEmphasis
glszm_SmallAreaLowGrayLevelEmphasis
ngtdm_Coarseness
glcm_ClusterShade
firstorder_Skewness
firstorder_Mean
firstorder_Median
Features
0.0
0.2
0.4
0.6
0.8
1.0
ICC


Wavelet_HLL
glszm_ZoneVariance
glszm_LargeAreaEmphasis
glszm_LargeAreaHighGrayLevelEmphasis
gldm_GrayLevelNonUniformity
glrlm_GrayLevelNonUniformity
glszm_LargeAreaLowGrayLevelEmphasis
glrlm_RunLengthNonUniformity
glrlm_RunEntropy
glcm_DifferenceEntropy
glrlm_GrayLevelNonUniformityNormalized
glszm_GrayLevelNonUniformityNormalized
glszm_GrayLevelNonUniformity
firstorder_Entropy
firstorder_Uniformity
glszm_GrayLevelVariance
firstorder_RootMeanSquared
glcm_InverseVariance
firstorder_Maximum
firstorder_10Percentile
firstorder_MeanAbsoluteDeviation
glcm_SumEntropy
ngtdm_Complexity
gldm_SmallDependenceEmphasis
glcm_DifferenceAverage
glcm_Id
glszm_ZonePercentage
gldm_DependenceNonUniformity
glcm_Idm
firstorder_RobustMeanAbsoluteDeviation
glcm_DifferenceVariance
glrlm_GrayLevelVariance
gldm_GrayLevelVariance
firstorder_Variance
glcm_Contrast
glszm_SizeZoneNonUniformityNormalized
glcm_JointEntropy
firstorder_InterquartileRange
glszm_SmallAreaEmphasis
firstorder_Range
glrlm_ShortRunEmphasis
glcm_SumSquares
glrlm_RunLengthNonUniformityNormalized
glszm_SizeZoneNonUniformity
ngtdm_Busyness
glrlm_RunPercentage
glcm_ClusterTendency
gldm_DependenceNonUniformityNormalized
glrlm_LongRunEmphasis
firstorder_Kurtosis
glcm_JointEnergy
firstorder_Energy
firstorder_TotalEnergy
gldm_LargeDependenceHighGrayLevelEmphasis
glrlm_RunVariance
glcm_MaximumProbability
gldm_LargeDependenceEmphasis
glcm_Imc2
glszm_ZoneEntropy
firstorder_Minimum
firstorder_90Percentile
glcm_ClusterProminence
ngtdm_Contrast
gldm_SmallDependenceHighGrayLevelEmphasis
gldm_DependenceEntropy
glcm_JointAverage
gldm_DependenceVariance
glszm_SmallAreaHighGrayLevelEmphasis
glrlm_LongRunHighGrayLevelEmphasis
ngtdm_Strength
glszm_HighGrayLevelZoneEmphasis
glrlm_HighGrayLevelRunEmphasis
gldm_HighGrayLevelEmphasis
glrlm_ShortRunHighGrayLevelEmphasis
glcm_Imc1
glcm_Idn
glcm_Autocorrelation
gldm_LargeDependenceLowGrayLevelEmphasis
ngtdm_Coarseness
firstorder_Mean
glcm_Correlation
firstorder_Median
glcm_Idmn
glszm_LowGrayLevelZoneEmphasis
glszm_SmallAreaLowGrayLevelEmphasis
glrlm_LongRunLowGrayLevelEmphasis
gldm_SmallDependenceLowGrayLevelEmphasis
glrlm_LowGrayLevelRunEmphasis
glrlm_ShortRunLowGrayLevelEmphasis
gldm_LowGrayLevelEmphasis
firstorder_Skewness
glcm_ClusterShade
Features
0.0
0.2
0.4
0.6
0.8
1.0
ICC

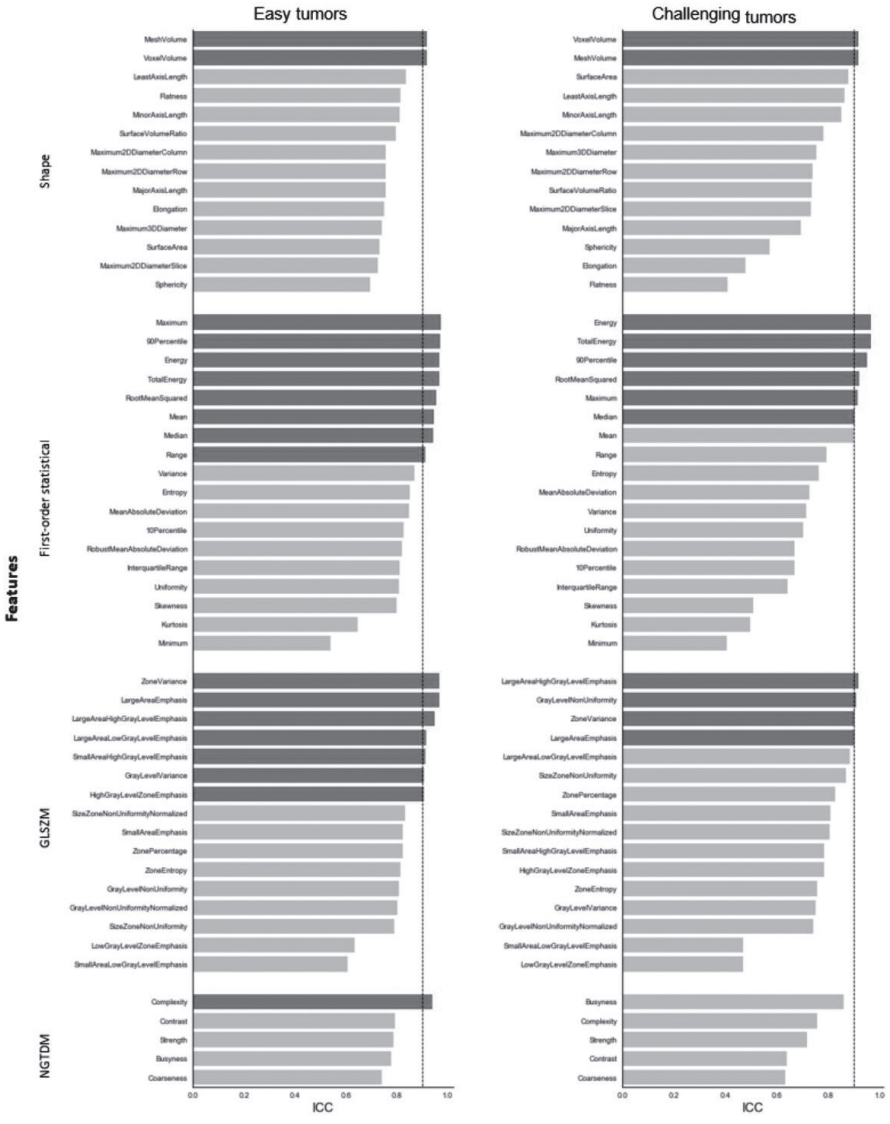

Wavelet_LHH
glcm_ClusterProminence
glcm_DifferenceVariance
glrlm_GrayLevelVariance
ngtdm_Complexity
firstorder_Variance
gldm_GrayLevelVariance
glszm_GrayLevelVariance
glcm_Contrast
glszm_LargeAreaHighGrayLevelEmphasis
glcm_SumSquares
glcm_ClusterTendency
glrlm_RunEntropy
firstorder_RootMeanSquared
gldm_SmallDependenceHighGrayLevelEmphasis
firstorder_MeanAbsoluteDeviation
glcm_DifferenceAverage
glcm_DifferenceEntropy
firstorder_90Percentile
firstorder_10Percentile
glcm_JointEntropy
glszm_ZoneVariance
glszm_LargeAreaEmphasis
firstorder_Entropy
glszm_SmallAreaHighGrayLevelEmphasis
firstorder_RobustMeanAbsoluteDeviation
glszm_GrayLevelNonUniformityNormalized
glszm_HighGrayLevelZoneEmphasis
glrlm_ShortRunHighGrayLevelEmphasis
firstorder_Range
gldm_GrayLevelNonUniformity
glcm_SumEntropy
glrlm_HighGrayLevelRunEmphasis
gldm_HighGrayLevelEmphasis
firstorder_InterquartileRange
glcm_Autocorrelation
glcm_InverseVariance
glrlm_GrayLevelNonUniformity
glrlm_GrayLevelNonUniformityNormalized
glrlm_LongRunHighGrayLevelEmphasis
firstorder_Minimum
glcm_JointAverage
firstorder_Maximum
glcm_Id
gldm_LargeDependenceHighGrayLevelEmphasis
glcm_Idm
firstorder_Uniformity
gldm_LargeDependenceLowGrayLevelEmphasis
glszm_GrayLevelNonUniformity
glrlm_RunLengthNonUniformity
ngtdm_Busyness
glcm_Correlation
firstorder_Energy
firstorder_TotalEnergy
glszm_SizeZoneNonUniformity
glcm_MaximumProbability
gldm_DependenceNonUniformity
glszm_LargeAreaLowGrayLevelEmphasis
gldm_SmallDependenceEmphasis
gldm_DependenceNonUniformityNormalized
glcm_JointEnergy
glrlm_RunLengthNonUniformityNormalized
glrlm_ShortRunEmphasis
firstorder_Kurtosis
glszm_ZonePercentage
glrlm_RunPercentage
glszm_ZoneEntropy
gldm_DependenceEntropy
gldm_LargeDependenceEmphasis
gldm_DependenceVariance
glrlm_LongRunEmphasis
glrlm_LongRunLowGrayLevelEmphasis
glcm_ClusterShade
glrlm_RunVariance
ngtdm_Contrast
gldm_LowGrayLevelEmphasis
glszm_SizeZoneNonUniformityNormalized
glcm_Idn
glrlm_LowGrayLevelRunEmphasis
glcm_Imc2
glszm_SmallAreaEmphasis
glrlm_ShortRunLowGrayLevelEmphasis
glszm_LowGrayLevelZoneEmphasis
glcm_Idmn
ngtdm_Coarseness
ngtdm_Strength
glszm_SmallAreaLowGrayLevelEmphasis
firstorder_Skewness
glcm_Imc1
gldm_SmallDependenceLowGrayLevelEmphasis
firstorder_Mean
firstorder_Median
Features
0.0
0.2
0.4
0.6
0.8
1.0
ICC
Wavelet_LHL
glcm_ClusterProminence
glszm_ZoneVariance
glszm_LargeAreaEmphasis
glszm_LargeAreaHighGrayLevelEmphasis
glszm_LargeAreaLowGrayLevelEmphasis
glszm_GrayLevelVariance
glcm_SumSquares
glcm_ClusterTendency
glcm_DifferenceVariance
glrlm_GrayLevelVariance
gldm_GrayLevelVariance
firstorder_Variance
ngtdm_Complexity
glcm_Contrast
gldm_GrayLevelNonUniformity
glrlm_GrayLevelNonUniformity
glrlm_RunEntropy
glrlm_RunLengthNonUniformity
glcm_JointEntropy
glcm_DifferenceEntropy
glcm_SumEntropy
glszm_GrayLevelNonUniformityNormalized
firstorder_RootMeanSquared
firstorder_MeanAbsoluteDeviation
glcm_DifferenceAverage
firstorder_Entropy
glszm_GrayLevelNonUniformity
ngtdm_Busyness
gldm_DependenceNonUniformity
glszm_SizeZoneNonUniformity
firstorder_InterquartileRange
firstorder_RobustMeanAbsoluteDeviation
firstorder_10Percentile
glrlm_GrayLevelNonUniformityNormalized
glcm_MaximumProbability
firstorder_Energy
firstorder_TotalEnergy
firstorder_Uniformity
glcm_JointEnergy
firstorder_90Percentile
gldm_SmallDependenceHighGrayLevelEmphasis
glcm_Id
firstorder_Range
glcm_InverseVariance
glcm_Idm
firstorder_Minimum
glrlm_ShortRunEmphasis
glrlm_LongRunEmphasis
glcm_JointAverage
firstorder_Maximum
glrlm_RunPercentage
glrlm_RunLengthNonUniformityNormalized
gldm_SmallDependenceEmphasis
glrlm_RunVariance
glszm_SmallAreaHighGrayLevelEmphasis
glszm_ZonePercentage
gldm_LargeDependenceEmphasis
gldm_DependenceNonUniformityNormalized
glszm_ZoneEntropy
glszm_SizeZoneNonUniformityNormalized
glszm_SmallAreaEmphasis
glrlm_ShortRunHighGrayLevelEmphasis
glszm_HighGrayLevelZoneEmphasis
glcm_Imc2
glrlm_HighGrayLevelRunEmphasis
gldm_HighGrayLevelEmphasis
gldm_DependenceEntropy
ngtdm_Strength
glcm_Autocorrelation
glrlm_LongRunHighGrayLevelEmphasis
firstorder_Skewness
gldm_DependenceVariance
ngtdm_Coarseness
gldm_LargeDependenceHighGrayLevelEmphasis
glcm_Imc1
firstorder_Kurtosis
ngtdm_Contrast
glcm_Idn
glcm_Idmn
glcm_ClusterShade
firstorder_Mean
glszm_SmallAreaLowGrayLevelEmphasis
firstorder_Median
gldm_SmallDependenceLowGrayLevelEmphasis
glszm_LowGrayLevelZoneEmphasis
glcm_Correlation
glrlm_ShortRunLowGrayLevelEmphasis
glrlm_LowGrayLevelRunEmphasis
gldm_LowGrayLevelEmphasis
glrlm_LongRunLowGrayLevelEmphasis
gldm_LargeDependenceLowGrayLevelEmphasis
Features
0.0
0.2
0.4
0.6
0.8
1.0
ICC

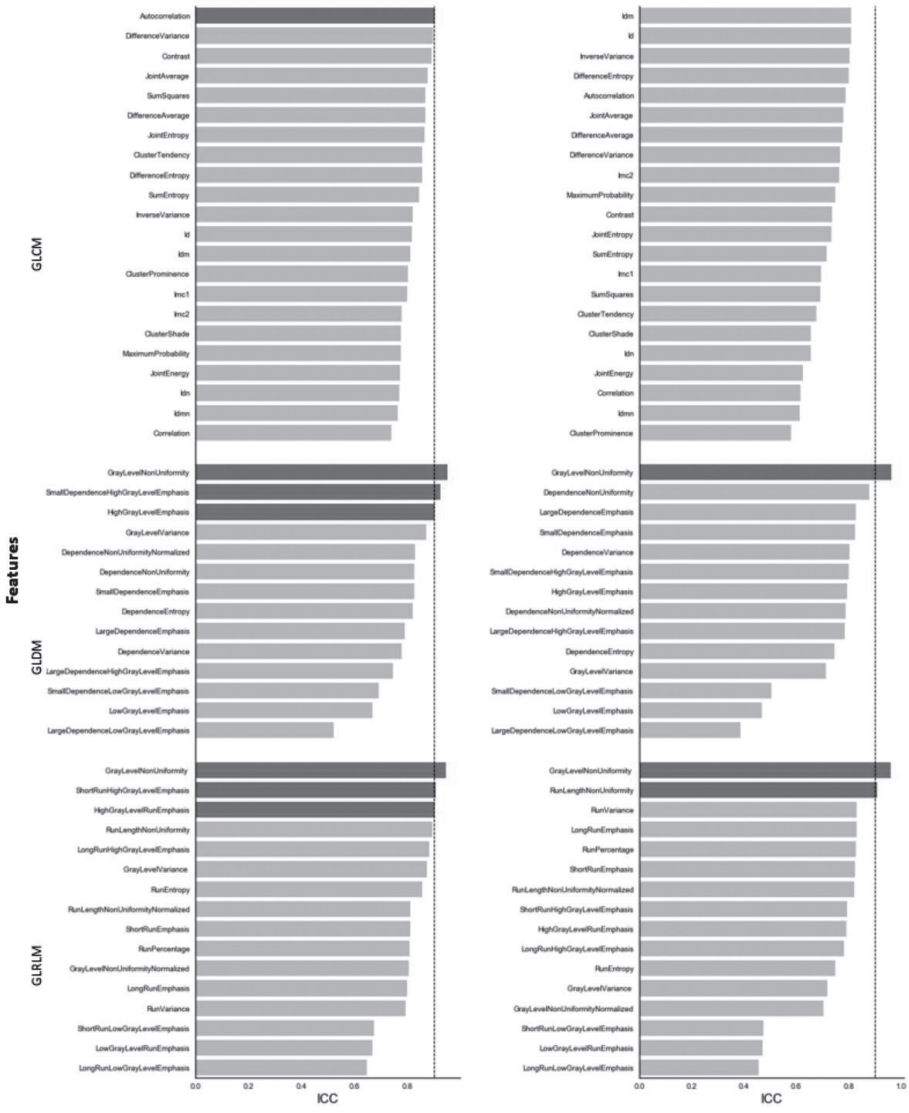

Wavelet_LLH
glszm_LargeAreaHighGrayLevelEmphasis
glszm_ZoneVariance
glszm_LargeAreaEmphasis
ngtdm_Complexity
gldm_GrayLevelNonUniformity
glrlm_GrayLevelNonUniformity
firstorder_Range
glszm_GrayLevelNonUniformityNormalized
glrlm_RunEntropy
firstorder_Maximum
gldm_SmallDependenceHighGrayLevelEmphasis
glcm_DifferenceEntropy
glszm_LargeAreaLowGrayLevelEmphasis
glszm_GrayLevelVariance
glcm_Correlation
firstorder_Entropy
glszm_SmallAreaHighGrayLevelEmphasis
glcm_ClusterProminence
glrlm_RunLengthNonUniformity
glcm_DifferenceAverage
glrlm_GrayLevelNonUniformityNormalized
glcm_DifferenceVariance
glrlm_GrayLevelVariance
glrlm_ShortRunHighGrayLevelEmphasis
glcm_SumEntropy
gldm_GrayLevelVariance
firstorder_Variance
gldm_HighGrayLevelEmphasis
glrlm_HighGrayLevelRunEmphasis
glszm_HighGrayLevelZoneEmphasis
glcm_JointEntropy
glcm_Contrast
firstorder_Minimum
glcm_SumSquares
firstorder_RootMeanSquared
glcm_Autocorrelation
glcm_ClusterTendency
firstorder_10Percentile
firstorder_Uniformity
glrlm_LongRunHighGrayLevelEmphasis
firstorder_MeanAbsoluteDeviation
glcm_JointAverage
glcm_InverseVariance
glcm_Id
glszm_GrayLevelNonUniformity
glcm_Idm
glszm_ZonePercentage
gldm_SmallDependenceEmphasis
gldm_DependenceNonUniformity
firstorder_Kurtosis
firstorder_RobustMeanAbsoluteDeviation
firstorder_Energy
firstorder_TotalEnergy
glszm_SizeZoneNonUniformity
gldm_LargeDependenceLowGrayLevelEmphasis
glrlm_RunLengthNonUniformityNormalized
firstorder_InterquartileRange
glrlm_ShortRunEmphasis
gldm_LargeDependenceHighGrayLevelEmphasis
gldm_DependenceNonUniformityNormalized
glrlm_RunPercentage
glszm_SmallAreaEmphasis
glcm_JointEnergy
glszm_SizeZoneNonUniformityNormalized
glcm_MaximumProbability
glrlm_LongRunEmphasis
firstorder_90Percentile
glrlm_RunVariance
gldm_LargeDependenceEmphasis
glszm_ZoneEntropy
glcm_Imc2
ngtdm_Busyness
glszm_SmallAreaLowGrayLevelEmphasis
glszm_LowGrayLevelZoneEmphasis
gldm_DependenceEntropy
glcm_ClusterShade
glrlm_LongRunLowGrayLevelEmphasis
gldm_DependenceVariance
glcm_Idn
glrlm_LowGrayLevelRunEmphasis
gldm_LowGrayLevelEmphasis
glrlm_ShortRunLowGrayLevelEmphasis
gldm_SmallDependenceLowGrayLevelEmphasis
ngtdm_Contrast
glcm_Idmn
ngtdm_Strength
firstorder_Skewness
glcm_Imc1
ngtdm_Coarseness
firstorder_Median
firstorder_Mean
Features
0.0
0.2
0.4
0.6
0.8
1.0
ICC
Wavelet_LLL
firstorder_Maximum
firstorder_Energy
firstorder_TotalEnergy
firstorder_90Percentile
gldm_GrayLevelNonUniformity
glrlm_GrayLevelNonUniformity
firstorder_RootMeanSquared
firstorder_Mean
firstorder_Median
glszm_GrayLevelNonUniformity
glrlm_RunLengthNonUniformity
ngtdm_Complexity
firstorder_Range
gldm_SmallDependenceHighGrayLevelEmphasis
glszm_SmallAreaHighGrayLevelEmphasis
glszm_HighGrayLevelZoneEmphasis
glrlm_ShortRunHighGrayLevelEmphasis
glrlm_HighGrayLevelRunEmphasis
gldm_HighGrayLevelEmphasis
glszm_ZoneVariance
glcm_Autocorrelation
glszm_LargeAreaEmphasis
glcm_DifferenceVariance
glrlm_LongRunHighGrayLevelEmphasis
gldm_DependenceNonUniformity
glcm_JointAverage
glszm_SizeZoneNonUniformity
glcm_DifferenceEntropy
glszm_GrayLevelVariance
glcm_Contrast
gldm_LargeDependenceEmphasis
glcm_DifferenceAverage
glrlm_RunVariance
glrlm_LongRunEmphasis
glrlm_ShortRunEmphasis
glrlm_GrayLevelVariance
glrlm_RunPercentage
ngtdm_Busyness
firstorder_Entropy
gldm_GrayLevelVariance
firstorder_Variance
glrlm_RunLengthNonUniformityNormalized
glszm_SmallAreaEmphasis
glszm_ZonePercentage
gldm_DependenceVariance
glcm_Idm
glcm_Id
glrlm_RunEntropy
glcm_InverseVariance
glszm_SizeZoneNonUniformityNormalized
glcm_Imc2
gldm_SmallDependenceEmphasis
glcm_SumSquares
firstorder_MeanAbsoluteDeviation
glcm_ClusterTendency
gldm_DependenceNonUniformityNormalized
firstorder_10Percentile
glcm_SumEntropy
glszm_ZoneEntropy
ngtdm_Contrast
glcm_JointEntropy
firstorder_Uniformity
glrlm_GrayLevelNonUniformityNormalized
firstorder_RobustMeanAbsoluteDeviation
gldm_LargeDependenceHighGrayLevelEmphasis
gldm_DependenceEntropy
firstorder_InterquartileRange
glcm_ClusterShade
glszm_GrayLevelNonUniformityNormalized
firstorder_Skewness
glszm_LargeAreaHighGrayLevelEmphasis
glcm_ClusterProminence
glcm_Imc1
glcm_Idn
ngtdm_Strength
glcm_MaximumProbability
glcm_Idmn
ngtdm_Coarseness
glcm_Correlation
gldm_SmallDependenceLowGrayLevelEmphasis
glcm_JointEnergy
glrlm_ShortRunLowGrayLevelEmphasis
gldm_LowGrayLevelEmphasis
glrlm_LowGrayLevelRunEmphasis
glrlm_LongRunLowGrayLevelEmphasis
glszm_LowGrayLevelZoneEmphasis
glszm_SmallAreaLowGrayLevelEmphasis
firstorder_Kurtosis
firstorder_Minimum
glszm_LargeAreaLowGrayLevelEmphasis
gldm_LargeDependenceLowGrayLevelEmphasis
Features
0.0
0.2
0.4
0.6
0.8
1.0
ICC

Supplementary material 3. ICC values of all unfiltered RadiomiX features with robust features (ICC > 0.9) shown in dark gray for all 'easy tumors' and 'challenging tumors'.


Features
NGTDM
GLCM
GLDZM
GLRLM
GLSZM
Fractal
complexity
contrast
strength
coarseness
busyness
homogeneity2
homogeneity1
inverseVar
diffVar
dissimilar
diffAvrg
diffEntro
contrast
sumSquares
autocorr
entrop2
sumVar
clusTend
infoCorr1
sumEntro
sumAvg
average
maxProb
infoCorr2
clusProm
invDiffNorm
energy
invDiffMomNor
maxCorr
clusShade
correl1
SZV
LDE
ZP
IV
HILDE
HIE
DZE
SDE
DZNN
HISDE
INN
IN
DZN
LISDE
LIE
LILDE
RLV
LRE
RP
RLNN
SRE
GLN
SRHGE
GLV
HGRE
RLN
LRHGE
RE
GLNN
SRLGE
LGRE
LRLGE
SZV
LAE
HILAE
ZP
IV
LILAE
HISAE
HIE
SZNN
SAE
ZE
INN
SZN
IN
LIE
LISAE
lacunarity
sd
average
0.0
0.2
0.4
0.6
0.8
1.0
ICC
busyness
strength
coarseness
complexity
contrast
inverseVar
homogeneity2
homogeneity1
diffEntro
diffAvrg
dissimilar
diffVar
contrast
maxProb
infoCorr1
infoCorr2
autocorr
entrop2
average
sumAvg
sumEntro
sumSquares
invDiffNorm
clusShade
clusTend
sumVar
correl1
energy
invDiffMomNor
clusProm
maxCorr
IN
ZP
DZN
HILDE
DZE
IV
SDE
DZNN
HIE
INN
LDE
DZV
HISDE
LISDE
LIE
LILDE
GLN
RLV
RLN
RP
LRE
RLNN
SRE
SRHGE
HGRE
LRHGE
RE
GLV
GLNN
SRLGE
LGRE
LRLGE
HILAE
SZV
LAE
IN
ZP
SZN
SZNN
SAE
LILAE
IV
ZE
HISAE
HIE
INN
LISAE
LIE
sd
lacunarity
average
0.0
0.2
0.4
0.6
0.8
1.0
ICC

Supplementary material 4. ICC values of all unfiltered Pyradiomics features with robust features (ICC > 0.9) shown in dark gray for all 'easy tumors' and 'challenging tumors'.




Features
GLCM
GLDM
GLRLM
Autocorrelation
DifferenceVariance
Contrast
JointAverage
SumSquares
DifferenceAverage
JointEntropy
ClusterTendency
DifferenceEntropy
SumEntropy
InverseVariance
Id
Idm
ClusterProminence
Imc1
Imc2
ClusterShade
MaximumProbability
JointEnergy
Idn
Idmn
Correlation
GrayLevelNonUniformity
SmallDependenceHighGrayLevelEmphasis
HighGrayLevelEmphasis
GrayLevelVariance
DependenceNonUniformityNormalized
DependenceNonUniformity
SmallDependenceEmphasis
DependenceEntropy
LargeDependenceEmphasis
DependenceVariance
LargeDependenceHighGrayLevelEmphasis
SmallDependenceLowGrayLevelEmphasis
LowGrayLevelEmphasis
LargeDependenceLowGrayLevelEmphasis
GrayLevelNonUniformity
ShortRunHighGrayLevelEmphasis
HighGrayLevelRunEmphasis
RunLengthNonUniformity
LongRunHighGrayLevelEmphasis
GrayLevelVariance
RunEntropy
RunLengthNonUniformityNormalized
ShortRunEmphasis
RunPercentage
GrayLevelNonUniformityNormalized
LongRunEmphasis
RunVariance
ShortRunLowGrayLevelEmphasis
LowGrayLevelRunEmphasis
LongRunLowGrayLevelEmphasis
0.0
0.2
0.4
0.6
0.8
ICC
Idm
Id
InverseVariance
DifferenceEntropy
Autocorrelation
JointAverage
DifferenceAverage
DifferenceVariance
Imc2
MaximumProbability
Contrast
JointEntropy
SumEntropy
Imc1
SumSquares
ClusterTendency
ClusterShade
Idn
JointEnergy
Correlation
Idmn
ClusterProminence
GrayLevelNonUniformity
DependenceNonUniformity
LargeDependenceEmphasis
SmallDependenceEmphasis
DependenceVariance
SmallDependenceHighGrayLevelEmphasis
HighGrayLevelEmphasis
DependenceNonUniformityNormalized
LargeDependenceHighGrayLevelEmphasis
DependenceEntropy
GrayLevelVariance
SmallDependenceLowGrayLevelEmphasis
LowGrayLevelEmphasis
LargeDependenceLowGrayLevelEmphasis
GrayLevelNonUniformity
RunLengthNonUniformity
RunVariance
LongRunEmphasis
RunPercentage
ShortRunEmphasis
RunLengthNonUniformityNormalized
ShortRunHighGrayLevelEmphasis
HighGrayLevelRunEmphasis
LongRunHighGrayLevelEmphasis
RunEntropy
GrayLevelVariance
GrayLevelNonUniformityNormalized
ShortRunLowGrayLevelEmphasis
LowGrayLevelRunEmphasis
LongRunLowGrayLevelEmphasis
0.0
0.2
0.4
0.6
0.8
1.0
ICC

References

1. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990).* 2012;48(4):441-446.
2. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006.
3. Gillies R, Kinahan P, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology.* 2016;278(2):563-577.
4. Bogowicz M, Vuong D, Huellner MW, et al. CT radiomics and PET radiomics: ready for clinical implementation? *Q J Nucl Med Mol Imaging.* 2019.
5. Davnall F, Yip CS, Ljungqvist G, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging.* 2012;3(6):573-589.
6. Grossmann P, Stringfield O, El-Hachem N, et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife.* 2017;6.
7. Ibrahim A, Vallières M, Woodruff H, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. Paper presented at: Seminars in Nuclear Medicine2019.
8. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14(12):749-762.
9. Walsh S, de Jong EE, van Timmeren JE, et al. Decision support systems in oncology. *JCO clinical cancer informatics.* 2019;3:1-9.
10. Milenkovic J, Dalmis MU, Zgajnar J, Platel B. Textural analysis of early-phase spatiotemporal changes in contrast enhancement of breast lesions imaged with an ultrafast DCE-MRI protocol. *Med Phys.* 2017;44(9):4652-4664.
11. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. *NPJ Breast Cancer.* 2017;3:43.
12. Liu Z, Li Z, Qu J, et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res.* 2019.
13. Xiong Q, Zhou X, Liu Z, et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy. *Clin Transl Oncol.* 2019.
14. Cain EH, Saha A, Harowicz MR, Marks JR, Marcom PK, Mazurowski MA. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast cancer research and treatment.* 2018.
15. Waugh SA, Purdie CA, Jordan LB, et al. Magnetic resonance imaging texture analysis classification of primary breast cancer. *Eur Radiol.* 2016;26(2):322-330.
16. Monti S, Aiello M, Incoronato M, et al. DCE-MRI Pharmacokinetic-Based Phenotyping of Invasive Ductal Carcinoma: A Radiomic Study for Prediction of Histological Outcomes. *Contrast Media Mol Imaging.* 2018;2018:5076269.
17. Cui X, Wang N, Zhao Y, et al. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer using Radiomics Features of DCE-MRI. *Sci Rep.* 2019;9(1):2240.

18. Yang J, Wang T, Yang L, et al. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer Using Mammography-Based Radiomics Method. *Sci Rep.* 2019;9(1):4429.
19. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2019;135:107-114.
20. Qiu Q, Duan J, Duan Z, et al. Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: impact of tumor segmentation variability. *Quant Imaging Med Surg.* 2019;9(3):453-464.
21. Belli ML, Mori M, Broggi S, et al. Quantifying the robustness of [(18)F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. *Phys Med.* 2018;49:105-111.
22. Pavic M, Bogowicz M, Wurms X, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol.* 2018;57(8):1070-1074.
23. Tixier F, Um H, Young RJ, Veeraraghavan H. Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features. *Med Phys.* 2019.
24. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. *Phys Med.* 2019;61:44-51.
25. Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp.* 2018;2(1):36.
26. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int J Radiat Oncol Biol Phys.* 2018;102(4):1143-1158.
27. Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol.* 2017;90(1070):20160665.
28. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol.* 2016;61(13):R150-R166.
29. Zwanenburg A LS, Valli`eres M, L¨ock S. Image biomarker standardisation initiative. . *arXiv preprint arXiv:161207003.*
30. Polan DF, Brady SL, Kaufman RA. Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study. *Phys Med Biol.* 2016;61(17):6553-6569.
31. Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *J Med Phys.* 2008;33(4):136-140.
32. Beresford MJ, Padhani AR, Taylor NJ, et al. Inter- and intraobserver variability in the evaluation of dynamic breast cancer MRI. *J Magn Reson Imaging.* 2006;24(6):1316-1325.
33. Saha A, Grimm LJ, Harowicz M, et al. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. *Med Phys.* 2016;43(8):4558.
34. Lobbes MB, Prevos R, Smidt M, et al. The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: a systematic review. *Insights Imaging.* 2013;4(2):163-175.
35. Houssami N, Turner R, Morrow M. Preoperative magnetic resonance imaging in breast cancer: meta-analysis of surgical outcomes. *Annals of surgery.* 2013;257(2):249-255.

36. Woolf DK, Padhani AR, Taylor NJ, et al. Assessing response in breast cancer with dynamic contrast-enhanced magnetic resonance imaging: are signal intensity-time curves adequate? *Breast Cancer Res Treat.* 2014;147(2):335-343.
37. Cardoso F, Kyriakides S, Ohno S, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* 2019.
38. El Khouli RH, Macura KJ, Jacobs MA, et al. Dynamic contrast-enhanced MRI of the breast: quantitative method for kinetic curve type assessment. *AJR Am J Roentgenol.* 2009;193(4):W295-300.
39. Pyradiomics feature description. https://pyradiomics.readthedocs.io/en/latest/features.html.
40. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017;77(21):e104-e107.
41. Ramesh A, Kambhampati C, Monson JR, Drew P. Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England.* 2004;86(5):334.
42. Measures of the Amount of Ecologic Association Between Species. *Ecology.* 1945;26(3):297-302.
43. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE transactions on medical imaging.* 1994;13(4):716-724.
44. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med.* 2016;15(2):155-163.
45. Racine JS. RStudio: a platform-independent IDE for R and Sweave. *Journal of Applied Econometrics.* 2012;27(1):167-172.
46. Gamer M, Lemon J, Fellows I, Sing P. Various Coefficients of Interrater Reliability and Agreement. *IRR: R package version 084.* 2012.
47. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography.* 2016;2(4):361-365.
48. Saha A, Harowicz MR, Mazurowski MA. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. *Med Phys.* 2018;45(7):3076-3085.
49. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv preprint arXiv:161207003.* 2016.
50. Hong J, Park BY, Lee MJ, Chung CS, Cha J, Park H. Two-step deep neural network for segmentation of deep white matter hyperintensities in migraineurs. *Comput Methods Programs Biomed.* 2019;183:105065.
51. Ghavami N, Hu Y, Gibson E, et al. Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. *Med Image Anal.* 2019;58:101558.
52. Kugelman J, Alonso-Caneiro D, Read SA, et al. Automatic choroidal segmentation in OCT images using supervised deep learning methods. *Sci Rep.* 2019;9(1):13298.
53. Abernethy AP, Etheredge LM, Ganz PA, et al. Rapid-learning system for cancer care. *Journal of Clinical Oncology.* 2010;28(27):4268.
54. Heye T, Merkle EM, Reiner CS, et al. Reproducibility of dynamic contrast-enhanced MR imaging. Part II. Comparison of intra-and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis. *Radiology.* 2013;266(3):812-821.

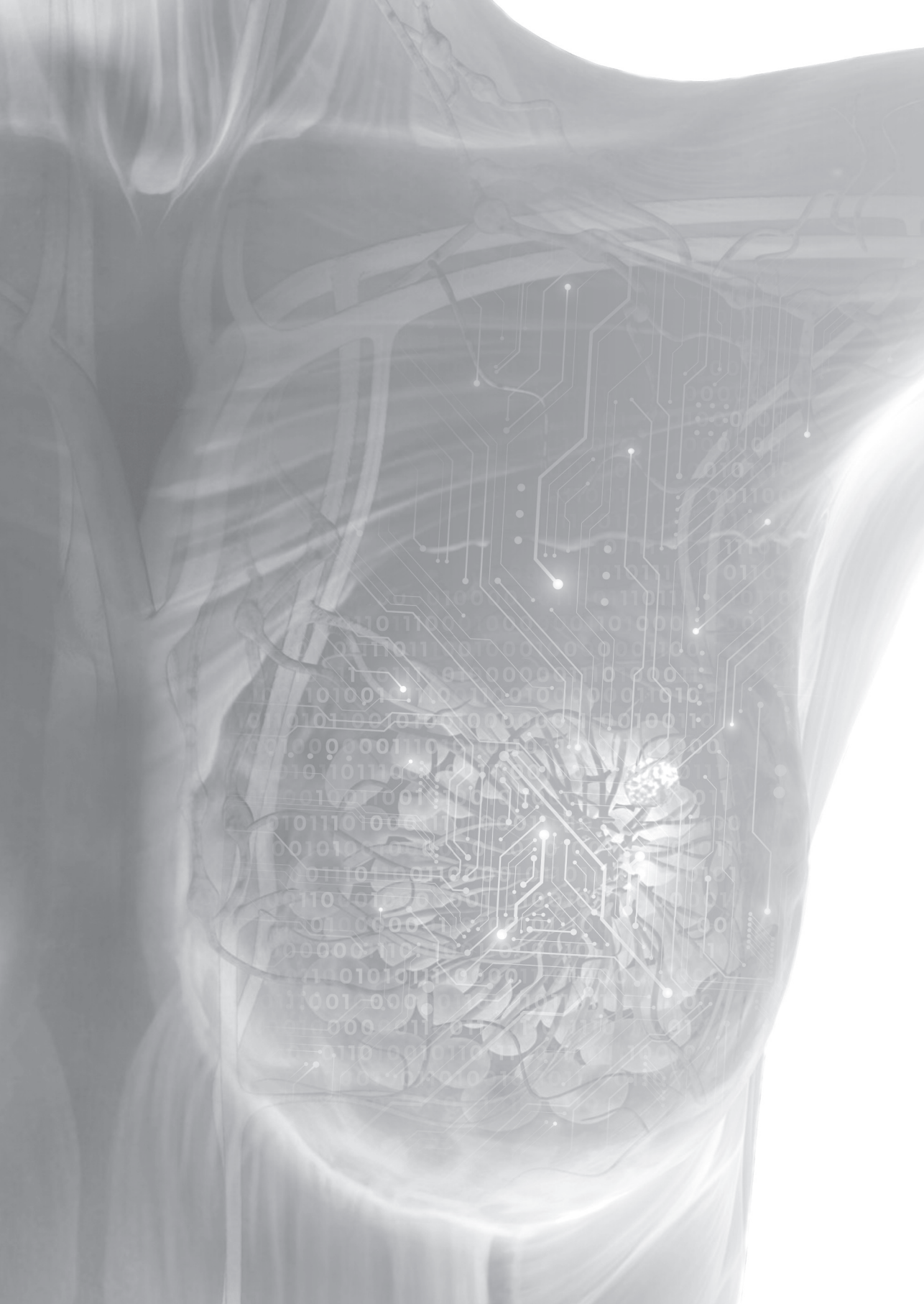
55. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Acad Radiol.* 2006;13(1):63-72.

56. Lin MQ, Chen JH, Wang XY, Chan SW, Chen SP, Su MY. Template-based automatic breast segmentation on MRI by excluding the chest region. *Med Phys Journal Translated Name Medical Physics.* 2013;40(12).

57. Thakran S, Chatterjee S, Singhal M, Gupta RK, Singh A. Automatic outer and inner breast tissue segmentation using multi-parametric MRI images of breast tumor patients. *PLoS One.* 2018;13(1):e0190348.

58. Fuller CD, Nijkamp J, Duppen JC, et al. Prospective randomized double-blind pilot study of site-specific consensus atlas implementation for rectal cancer target volume delineation in the cooperative group setting. *Int J Radiat Oncol Biol Phys.* 2011;79(2):481-489.

59. Mitchell DM, Perry L, Smith S, et al. Assessing the effect of a contouring protocol on postprostatectomy radiotherapy clinical target volumes and interphysician variation. *Int J Radiat Oncol Biol Phys.* 2009;75(4):990-993.



CHAPTER 8

Discussion and future perspectives

The conversion of medical images into quantitative, robust, and generalizable data that can contribute to clinical decision support systems, answering questions about diagnosis, response predictions, and prognoses is the goal of the application of radiomics. With the introduction of the radiomics framework by Lambin et al. [1] in 2012, it became possible to extract quantitative information from medical imaging that cannot be obtained through the visual assessment of the radiologist. Ultimately, these clinical decision support systems should contribute to the enhancement of personalized medicine. The use of radiomics has increased exponentially over the past decade, with the vast majority of published radiomics articles focusing on the oncology field [2]. However, the actual clinical applicability of radiomics in oncology is lagging. Radiomics is currently being investigated in varying imaging modalities for different tumor sites. In this thesis, we focused on the application and optimization of radiomics on magnetic resonance imaging (MRI) for predictive analysis in breast cancer treatment.

The first part of this thesis focused on the predictive power of MRI-based radiomics in breast cancer patients. We assessed the current state of tumor response prediction of MRI-based radiomics through a systematic review, followed by the development of multiple radiomics prediction models encompassing both the breast tumor and axillary lymph nodes.

During the development of the MRI-based radiomics models, several pitfalls in the radiomics workflow came to light. One of these pitfalls was the unknown variability of radiomic features for different MRI scanner acquisition and reconstruction parameters. It is imperative to investigate and quantify this previously unknown variability as the radiomic analysis needs quantitative data. Therefore, the second part of this thesis focused on two of these pitfalls to optimize the radiomics workflow used for breast MRI-radiomics studies. Here we examined the robustness of features with regard to variable manual tumor segmentation and repeatability of features when extracted from multiple MRI exams of eleven healthy volunteers in three test-retest settings.

Part I - MRI-based radiomics for prediction purposes in breast cancer patients

Response prediction using MRI-based radiomics

The percentage of breast cancer patients receiving neoadjuvant systemic therapy (NST) varies between 13-59%, depending on breast cancer subtype, with the highest rates in triple-negative and HER2-positive breast tumors [3,4]. The use of breast tumor response to NST for patient-tailored treatment adjustments has been studied for many years. For example, the addition of neo-adjuvant

trastuzumab has been shown to improve pathologic complete tumor response (pCR) rates and event-free survival in HER2-positive breast tumors [5,6]. Tumor response to NST has also changed surgical treatment over the years. Where we have moved from ablative surgery to breast-conserving surgery in a large subset of the breast cancer population, we are now about to forgo surgery in breast cancer patients who have a confirmed complete response after NST, as approximately 30% of all breast cancer patients treated with NST achieves a pCR. However, to date, it is not possible to predict a pathological complete tumor response with imaging performed before surgery. Even in combination with patient and tumor characteristics, the diagnostic accuracy of routinely used medical imaging is insufficient to predict pCR [7]. This is despite continuous improvements in imaging equipment, imaging techniques, and image quality and accuracy. For that reason, pathology examination after surgery, after completion of NST, is still the gold standard to determine treatment response.

More recently, the use of image-guided biopsies has been explored as a tool to confirm breast tumor pCR. While relatively small, single-center studies showed promising results [8,9], larger multicenter studies showed poor performance with false-negative rates of up to 50% [10-12]. As a follow-up to these studies, multivariate analyses were performed, supplementing the image-guided biopsy information with routine clinical imaging, and patient and tumor characteristics, resulting in a false-negative rate of 1.2% [13]. Research into radiomics features, that can be treated as biomarkers and contribute to these analyses, is an important area of research. Achieving an accurate pCR prediction may therefore mean that approximately 30% of breast cancer patients, surgery and/or adjuvant therapy can be safely omitted, thereby reducing the patients' complication risk and thus the risk of co-morbidities. Since MRI is the most widely used in the clinic and most accurate breast imaging for monitoring tumor response to NST, the majority of research on radiomics in breast cancer has focused on MRI-based radiomics.

To determine the current state of MRI-based radiomics for the prediction of tumor response to NST in breast cancer patients, a descriptive systemic review was conducted. A total of sixteen studies were analyzed describing 1736 patients, ranging from 35 to 414 patients per article. The overall Radiomics Quality Scores were considered as "poor" with a mean of 11% (range 0% – 41.2%), with the most recent articles receiving the highest score. Failure to perform external validation causes the greatest loss of quality. The study of van Timmeren et al. [14] showed that external validation is necessary to investigate the applicability and generalizability of a model based on retrospective multicenter data cohorts. Besides the lack of external validation data cohorts, further findings of the systematic review showed large methodological heterogeneity among the included articles. Due to these methodological differences, the (overall promising) results of the individual studies could not be compared.

In this thesis, we performed MRI-based radiomics analysis for two different prediction purposes. In chapter four we investigated the ability of the pretreatment MRI to predict tumor response to NST in breast cancer patients using radiomics analysis in a multicenter study. Initially, we hypothesized that radiomics models trained, tested, and validated on data from two independent data cohorts, could add information to the prediction of tumor response to NST, and that combined with clinical models could improve prediction accuracy. During data analyses, the sensitivity of radiomics features to variations in acquisition and reconstruction parameters, and variations in MRI scanners was established [15-18]. This insight, derived from recent radiomics publications and an emerging consensus in the field, made us realize that the initially formulated hypothesis was not feasible with the data available for this study. Therefore, we changed the hypothesis from the ability of radiomics models to accurately predict tumor response to NST when trained, tested, and validated on data from two independent cohorts acquired differently, in the inability of radiomics to accurately perform this prediction. Although, of course, it was not ruled out that the study could contribute to an accurate pCR prediction.

The collected MRI data varied in MRI vendor, magnetic field strength, and acquisition and reconstruction parameters. Due to the lack of reproducibility data (i.e. phantom data or test-retest data), we were not able to correct for these variabilities. A total of 322 tumors from 290 breast cancer patients were enrolled and analyzed using three different strategies. The AUC values of the radiomics, clinical, and combined models in the validation datasets of the three strategies had ranges of 0.52-0.57, 0.71-0.77, and 0.66 -0.74, respectively. These results show that the clinical models significantly outperformed the radiomics models, indicating that radiomics features in these scenarios did not have an added value to the clinical models developed. Besides, the radiomics features selected for the models and their performance differed with and within the three strategies. It was concluded that reproducibility studies are needed to determine the effects of different MRI scanners and different acquisition and reconstruction parameters on radiomics features before conclusions can be drawn from the results found in this study.

A previous study by Liu et al. [19] investigated the same topic and reached significantly better external validation results in the radiomics models with AUC values between 0.71 – 0.80. However, this study differed from our study by the use of multiparametric (multiple sequences) MRI-based radiomics, which appeared to improve the outcome. Noteworthy is that their external validation results consisted of data cohorts from three different hospitals using three different MRI vendors (GE, Philips, and Siemens). Furthermore, their magnetic field strengths varied with both 1.5 and 3.0 Tesla with different acquisition and reconstruction parameters, and the images were even much less extensively preprocessed compared to our

images. Consequently, we cannot explain how such heterogeneous data can achieve such high AUC values. Our request to exchange their data together with the developed models to reproduce results was not answered.

In chapter five, the ability of MRI-based radiomics analysis for the preoperative prediction of axillary lymph nodes metastasis in breast cancer patients was investigated. Accurate noninvasive preoperative prediction of axillary lymph node metastasis can assist in clinical decision-making and potentially spare breast cancer patients without axillary metastasis from axillary surgery. In this prospective study, radiomics features were extracted from segmented axillary lymph nodes of 75 breast cancer patients undergoing dedicated axillary MRI exams. In total, 511 axillary lymph nodes were segmented on the T2-weighted sequence of the dedicated axillary MR images. Of the included axillary lymph nodes, 36/511 (7%) were histologically confirmed as malignant axillary lymph nodes. For all developed models, each cohort split resulted in a different number of lymph nodes in the training cohorts and a different set of selected features. The performance of the radiomics models showed a wide range of AUC values between 0.48 – 0.89 and 0.37 – 0.99 for the training and validation models, respectively. Based on these results, it was not possible to develop a final radiomics prediction model.

Based on these limited results, two hypotheses can be formed. First, the variation in acquisition and reconstruction parameters significantly affects radiomics features, in such way that data is not readily interpretable and results are not reliable without further experimentation and analysis. Although this study is based on prospectively collected data allowing many parameters to be controlled, there was still variation in acquisition and reconstruction parameters depending on the patient. Second, MRI-based radiomics using dedicated axillary images and axillary lymph node segmentations does not have an added value in the prediction of axillary lymph node metastasis. To form a definitive judgment, the effects of variation in acquisition and reconstruction parameters on radiomics features must be determined. Therefore, it is necessary to perform phantom and reproducibility studies.

In summary, the developed radiomics models do not seem to contribute to either prediction question. However, both articles conclude that it is necessary to first investigate the sensitivity of MRI-based radiomics to different MRI scanners and varying acquisition and reconstruction parameters before it can be definitively established that radiomics models do not contribute to these specific prediction questions. For that reason, we investigated feature repeatability in the second part of this thesis.

Part II – Optimization in MRI-based radiomics

Optimizing MRI-based radiomics is related to improving the generalizability, comparability and reproducibility of radiomics studies. The current trend of published radiomics articles shows an emphasis on optimizing the methodological side of the radiomics workflow, rather than applying radiomics to diagnostic, predictive or prognostic problems [20-23]. The latest research mainly focuses on finding a solution for the reproducibility of radiomics features values in which different harmonization methods are investigated. This after several studies, including interobserver segmentation studies and test-retest studies, showed the sensitivity of radiomic features values to variations in scanners and acquisition parameters. Harmonization methods such as normalization, intensity harmonization, ComBat and deep learning methods have been investigated, showing varying results [24-26].

Furthermore, initiatives like the *Image Biomarker Standardization Initiative* (IBSI) have been established to optimize the radiomics workflow by standardizing the extraction of radiomics features by capturing formulas and names of the various radiomics features, standardizing the steps in the radiomics workflow, and verifying the radiomics softwares [27]. Furthermore, the TRIPOD guideline is useful in transparently reporting prognostic or diagnostic multivariable prediction models [28,29]. In addition, several articles extensively discuss the radiomics workflow and provide guidelines for performing radiomics analysis [30,31]. A recently published paper formulated this specifically for MRI-based radiomics in breast cancer [32]. The optimization of the radiomics workflow and reproducibility of results also benefits from the fact that journals publishing radiomics articles encourage scientists to publish their developed radiomics models (including datasets used) online. Specifically, in MRI-based radiomics, feature robustness is a hot topic that needs to be optimized before it is useful for radiomics analysis.

Feature robustness in MRI-based radiomics

The standard radiomics workflow consists of image acquisition, image pre-processing, region of interest (ROI) segmentation, feature extraction, feature selection, and feature analysis (e.g. model development). Every step contains obstacles that must be overcome to reach a radiomics workflow of sufficient quality in such a way that developed clinical decision support systems significantly contribute to current clinical practice. Besides the much-discussed necessity of identifying repeatable and reproducible features, most of these obstacles relate to the variability seen in all steps of the radiomics workflow. This variability prevents the radiomics community from achieving generalizable and comparable results that could drive advances in oncology clinical care.

Image acquisition is the first step of the radiomics workflow, therefore obstacles in this step must be addressed first to prevent them from continuing in subsequent steps. Here, the use of MRI scanners from different vendors, which in turn develop different types of MRI scanners with different magnetic field strengths and different scanning sequences, result in heterogeneous data. Moreover, the addition of contrast also causes heterogeneity, inter alia due to the biology of the patient but also due to the use of different types of contrast agents in different doses. Also, the refinement of the clinical protocols per hospital results in further heterogeneity. A recent phantom pelvic study by Bianchini et al. [33] showed that the reproducibility of radiomics features extracted from T2W images was affected by both the use of different MRI vendors as well as the difference in magnetic field strengths. Of both, the difference in MRI vendor had more effect on the reproducibility of the radiomics features than magnetic field strength. Solely, 4.6% (43/944) and 15.6% (147/944) of the radiomics features showed excellent reproducibility when extracted from images from different MRI vendors or MRI scanners with different magnetic field strengths, respectively.

The multi-center retrospective study in chapter four suffered from data heterogeneity discussed above. Ideally, as long as specific MRI test-retest data is missing, only data obtained from the same MRI scanner with an identical clinical protocol using fixed parameters should be used in future studies. However, this is impossible to implement, perhaps even undesirable in clinical practice. Reproducibility studies in terms of phantom and test-retest studies are needed to address this undesirable variability, especially for image acquisition. To date, most studies on the reproducibility of radiomics features have been based on phantom data, with most focusing on CT imaging. The downside of phantom studies is the lack of the human factor, including, but not limited to, patient positioning, body temperature, and respiration. Furthermore, results of phantom studies investigating repeatability and reproducibility seem to be overly optimistic compared to test-retest studies using human data[34].

Test-retest MRI studies have been performed in cervical cancer [23], prostate cancer [35,36], and glioblastoma [17,22,37-40]. In summary, these studies concluded that there is great variation in the repeatability of the different feature groups and that the repeatability of the features is highly sensitive to image pre-processing procedures. It is important to note that here again, feature comparability is not self-evident as not all studies use the same feature extraction software and if they do, there may still be a difference in the set of features extracted. In addition, the sensitivity of radiomics features to variations in the MRI acquisition is most likely tumor site specific and should therefore be examined per tumor site. For example, van Timmeren et al. [41] showed that over 80% of the radiomics features extracted from CT images of lung cancer patients had higher reproducibility scores compared to

the features extracted from CT images of rectal cancer patients. Contrary to this, a study by Peerlings et al. [17] identified 9.2% (122/1322) robust features on apparent diffusion coefficient maps in different tumor sites (lung, liver, and ovary), using different MR-systems and vendors with a 1.5T magnetic field strength. Given these contrasting results, it seems most obvious for now to first examine the robustness of features per tumor site, followed by a comparison between tumor sites.

In chapter six we performed a breast MRI test-retest study in eleven healthy volunteers who were scanned multiple times using three different test-retest settings on two different days using an identical clinical breast MRI protocol. For each scan, 91 radiomics features were extracted from the manual segmented right breast, before and after image pre-processing. The images without pre-processing produced the highest number of repeatable features for both the T1W sequence as the ADC maps with 15/91 (16.5%) and 8/91 (8.8%) repeatable features, respectively, using the concordance correlation coefficient (CCC) > 0.9 as cut-off value. In the T2W images, applying z-score normalization produced the highest number of repeatable features, 26/91 (28.6%). It was concluded that in addition to the MRI sequence, the image preprocessing also influences the repeatability of the radiomics features and that regardless of test-retest setting and scanning moment, only a limited number of features appeared to be repeatable. The results of this study can serve as a starting point for further research into reproducibility of breast MRI-based radiomics.

After image acquisition and pre-processing, ROI segmentation follows, a step prone to intra- and inter-observer variation [42,43]. Practice shows that segmentations are often still performed manually, as automatic segmentation software is not always available and semi-automatic segmentations often require a lot of manual adjustments. The difficulty of segmentations depends on the tumor site to be examined and the MRI sequence chosen. In chapter seven, we explored inter-observer manual segmentation variability in breast MRI exams and its effect on feature robustness, a study that has not been performed before. In 129 histologically confirmed breast tumors, segmentations were performed by four observers with different degrees of experience in MRI breast segmentation. The inter-observer variability evaluated by the volumetric Dice Similarity Coefficient resulted in a mean of 0.81. In comparison, cervical tumors segmented by multiple readers in the study by Fiset et al. [23] showed a mean Dice Similarity Coefficient above 0.9. The slightly lower value found in our study is probably due to the inclusion of some large and irregularly shaped breast tumors, which are more difficult to segment. Overall, a further increase of the Dice Similarity Coefficient could be reached by extensive and clear segmentation guidelines per tumor site or by the incorporation of an atlas-based approach. Furthermore, the feature robustness was assessed using the intraclass correlation coefficient (ICC > 0.9)

for features extracted by two commonly used radiomics software. The results showed that 41.6% robust features of the 1,328 radiomics features extracted using the RadiomiX toolbox (OncoRadiomics SA, Liege, Belgium), and 32.8% robust features of the 833 radiomics features extracted using the open-source Pyradiomics software. So, we concluded that more complex, challenging tumors result in less robust features. Finally, these identified robust features can be used as pre-selection after the selection of robust features in the first two steps of the radiomics workflow, image acquisition, and image pre-processing.

The results of the two studies performed in the second part of this thesis proved that radiomics features extracted from breast MRI exams are susceptible to variations induced by the radiomics workflow. It is the minority of features that appear to be robust to the variations studied. With regard to the variation in acquisition and reconstruction parameters across different MRI scanners, in addition to the test-retest study performed in this thesis, several research approaches have been performed to assess the robustness of radiomics features. A recent published review article by Atul Mali et al. [26] provided a comprehensive overview of studies investigating feature robustness using different harmonization methods. The different harmonization methods were divided into two domains; the 'image' domain where harmonization was performed on the complete raw or reconstructed image, and the 'feature' domain where harmonization was applied to the extracted radiomics features of non-harmonized images. The study concluded that to date there is no unequivocal method to solve the robustness problem in radiomics. Among other things, there was critical writing about the already widely used ComBat method, a method for estimating the effects of two batches with one technical difference, taking into account the effect of biological covariates on the radiomics features that have been harmonized. This is mainly due to the greater complexity of radiomics related to gene expression arrays, for which ComBat was once developed. Great potential is seen in the style transfer technique. This is a computerized technique that combines two images of different badges, the content image and the reference style image. The goal is that the final image retains the key elements of the content image, but appears 'painted' in the style of the reference image. Adding a generative adversarial network (GAN), a machine learning (ML) model in which two neural networks compete to create a more accurate prediction, would allow for further optimization. However, the deployment of GANs is complex and require a large amount of data. Ultimately, this review article concluded that large phantom and test-retest studies are needed to develop definitive harmonization methods that can be used for radiomics studies applied to clinical decision support systems.

A conclusion that also emerges from the second part of this thesis, in which the results of chapters six and seven provide a basis for further investigation into the robustness of radiomic features obtained from breast MRI exams. Options for further research are discussed in the future perspective.

Conclusions

- MRI-based tumor response prediction studies showed large methodological differences in the radiomics workflow which disturbs comparability.
- Dedicated axillary MRI-based radiomics with node-by-node analysis did not contribute to the prediction of axillary lymph node metastasis.
- MRI-based radiomics showed no contribution to the pretreatment prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients.
- However, for both prediction studies, results can only become definitive after the effect of different scanners and variation in acquisition and reconstruction parameters on feature values is known.
- Radiomics features extracted from T1W and T2W sequences and ADC maps from breast MRI exams showed a varying, limited number of repeatable features, wherein the effect of the preprocessing procedures for each sequence is different on the repeatability of features.
- Variations in manual MRI breast tumor segmentations affect feature values with less than half of the extracted RadiomiX and Pyradiomics features being robust to these variations.

Future perspective

The results of the MRI test-retest study revealed, per investigated MRI sequence, a set of repeatable features. These features are a starting point for the next step in unraveling robust MRI-based radiomics features.

After the identification of repeatable features in chapter six, one option for future research could be to focus on feature reproducibility. Reproducible features remain the same when imaged with different MRI scanners from different vendors, with different field strengths and different acquisition and reconstruction parameters. In breast MRI studies collected from breast cancer patients for both single-center and multicenter studies, these variations will always be present to a greater or lesser extent, therefore identifying reproducible features is a necessary next step. Feature reproducibility can be examined by changing acquisition parameters one by one while leaving the others fixed, followed by performing the same procedure on different MRI scanners from different vendors with different field strengths. This, to see if the results hold when MRI exams are manufactured on different MRI scanners.

Another option is to use the repeatable features found in this thesis on MRI exams of breast cancer patients, although this is only possible if the same scanning protocol is used on the same scanner. This seems practically feasible as the majority of breast cancer patients in the MUMC+ are scanned on the MRI scanner used in the MRI test-retest study in this thesis. Furthermore, the fixed scanning protocol studied in the test-retest article is based on the clinical breast scanning protocol of the MUMC+. In order to use the repeatable features from chapter six, it is first important to verify that the radiologist assesses MRI exams made with the fixed scanning protocol used in the test-retest study in chapter six the same as MRI exams made with the original breast scanning protocol. If so, repeatability in breast tumors can be investigated and then used for further radiomics analysis. The advantage of this is that time-consuming and expensive reproducibility studies are largely no longer necessary, but radiologists do have to agree on the use of fixed scanning protocols.

After the identification of repeatable and reproducible radiomics features, it would be interesting to develop prediction models based on specific breast cancer subtypes, as the percentage of breast cancer patients achieving a pCR is highly dependent on this, with the HER2-positive and triple negative breast cancer subtypes reaching the highest percentage.

Another data-driven method is deep learning (DL), a subfield of machine learning and inspired by the neural networks of the human brain. DL is already being widely applied in everyday life for a variety of tasks, including image and speech

recognition, and autonomous driving. However, clinical application of DL is still lacking. One of the limiting factors is that DL is seen as a 'black box', which hinders interpretability. Moreover, DL analysis requires large amounts of data to achieve reproducible results and the process of acquiring such datasets in the medical field is challenging itself. However, when the above-mentioned challenges are overcome, there is certainly a future for the DL algorithms in clinical use. It can be used as a standalone option to enhance personalized medicine or to contribute to the application of radiomics to enhance personalized medicine. Currently, DL is actively being used for applications working on automatic segmentation problems, which could eliminate segmentation variability and save a huge amount of time. Recently, our research group developed a fully automated CT lung tumor detection and 3D volumetric segmentation pipeline of non-small cell lung cancers, which can handle the differences in acquisition and reconstruction parameters of the CT scans. Now, a user-friendly web application is being made allowing for clinical use of the DL algorithm. The software has already achieved the CE mark as a medical device class 1 under the Medical Device Directives and soon will be used for the first clinical trials. Furthermore, research is now performed on applying deep learning for image normalization for multicenter studies, although more thorough research is needed to show its potential for use in the clinic [44].

Parallel to the above-mentioned future perspectives, further research can be done into harmonization methods. The review article of Mali et al. [26] extensively discussed the currently used harmonization methods as a solution for both standardizations of radiomic features and medical images. Although several methods show promising results they conclude that more research is needed to explore the boundaries of feature and image normalization methods. This means that future research must invest in large phantom and test-retest studies in order to arrive at harmonization methods that can eventually be used for radiomics studies that can be applied in the clinic.

References

1. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990).* 2012;48(4):441-446.
2. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol.* 2016;61(13): R150-R166.
3. Murphy BL, Boughey JC. ASO Author Reflections: Changes in Use of Neoadjuvant Chemotherapy Over Time-Highest Rates of Use Now in Triple-Negative and HER2+ Disease. *Annals of surgical oncology.* 2018.
4. Krystel-Whittemore M, Xu J, Brogi E, et al. Pathologic complete response rate according to HER2 detection methods in HER2-positive breast cancer treated with neoadjuvant systemic therapy. *Breast cancer research and treatment.* 2019;177(1):61-66.
5. Buzatto IP, Ribeiro-Silva A, Andrade JM, Carrara HH, Silveira WA, Tiezzi DG. Neoadjuvant chemotherapy with trastuzumab in HER2-positive breast cancer: pathologic complete response rate, predictive and prognostic factors. *Braz J Med Biol Res.* 2017;50(2):e5674.
6. Gianni L, Eiermann W, Semiglazov V, et al. Neoadjuvant and adjuvant trastuzumab in patients with HER2-positive locally advanced breast cancer (NOAH): follow-up of a randomised controlled superiority trial with a parallel HER2-negative cohort. *Lancet Oncol.* 2014;15(6):640-647.
7. Fowler AM, Mankoff DA, Joe BN. Imaging Neoadjuvant Therapy Response in Breast Cancer. *Radiology.* 2017;285(2):358-375.
8. Heil J, Schaefgen B, Sinn P, et al. Can a pathological complete response of breast cancer after neoadjuvant chemotherapy be diagnosed by minimal invasive biopsy? *European Journal of Cancer.* 2016;69:142-150.
9. Kuerer HM, Rauch GM, Krishnamurthy S, et al. A Clinical Feasibility Trial for Identification of Exceptional Responders in Whom Breast Cancer Surgery Can Be Eliminated Following Neoadjuvant Systemic Therapy. *Annals of surgery.* 2018;267(5):946-951.
10. van Loevezijn AA, van der Noordaa MEM, van Werkhoven ED, et al. Minimally Invasive Complete Response Assessment of the Breast After Neoadjuvant Systemic Therapy for Early Breast Cancer (MICRA trial): Interim Analysis of a Multicenter Observational Cohort Study. *Annals of surgical oncology.* 2020.
11. Basik M, Cecchini RS, De Los Santos JF, et al. Primary analysis of NRG-BR005, a phase II trial assessing accuracy of tumor bed biopsies in predicting pathologic complete response (pCR) in patients with clinical/radiological complete response after neoadjuvant chemotherapy (NCT) to explore the feasibility of breast-conserving treatment without surgery. *Cancer Res.* 2020;80(4).
12. Heil J, Pfob A, Sinn HP, et al. Diagnosing Pathologic Complete Response in the Breast After Neoadjuvant Systemic Treatment of Breast Cancer Patients by Minimal Invasive Biopsy: Oral Presentation at the San Antonio Breast Cancer Symposium on Friday, December 13, 2019, Program Number GS5-03. *Annals of surgery.* 2020.
13. Pfob A, Sidey-Gibbons C, Lee HB, et al. Identification of breast cancer patients with pathologic complete response in the breast after neoadjuvant systemic treatment by an intelligent vacuum-assisted biopsy. *European journal of cancer (Oxford, England : 1990).* 2021;143:134-146.
14. van Timmeren JE, Carvalho S, Leijenaar RTH, et al. Challenges and caveats of a multi-center retrospective radiomics study: an example of early treatment response assessment for NSCLC

patients using FDG-PET/CT radiomics. *PLoS One.* 2019;14(6):e0217536.

15. Rai R, Holloway LC, Brink C, et al. Multicenter evaluation of MRI-based radiomic features: A phantom study. *Med Phys.* 2020.
16. Dreher C, Kuder TA, Konig F, et al. Radiomics in diffusion data: a test-retest, inter- and intra-reader DWI phantom study. *Clin Radiol.* 2020;75(10):798 e713-798 e722.
17. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep.* 2019;9(1):4800.
18. Shur J, Blackledge M, D'Arcy J, et al. MRI texture feature repeatability and image acquisition factor robustness, a phantom study and in silico study. *Eur Radiol Exp.* 2021;5(1):2.
19. Liu Z, Li Z, Qu J, et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res.* 2019.
20. Scalco E, Belfatto A, Mastropietro A, et al. T2w-MRI signal normalization affects radiomics features reproducibility. *Med Phys.* 2020;47(4):1680-1691.
21. Pandey U, Saini J, Kumar M, Gupta R, Ingalhalikar M. Normative Baseline for Radiomics in Brain MRI: Evaluating the Robustness, Regional Variations, and Reproducibility on FLAIR Images. *J Magn Reson Imaging.* 2020.
22. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *J Appl Clin Med Phys.* 2020;21(1):179-190.
23. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2019;135:107-114.
24. Ibrahim A, Refaee T, Leijenaar RTH, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS One.* 2021;16(5):e0251147.
25. Ibrahim A, Refaee T, Primakov S, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers.* 2021;13(8).
26. Mali SA, Ibrahim A, Woodruff HC, et al. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J Pers Med.* 2021;11(9).
27. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology.* 2020.
28. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* 2015;13:1.
29. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
30. Rogers W, Thulasi Seetha S, Refaee TAG, et al. Radiomics: from qualitative to quantitative imaging. *Br J Radiol.* 2020;93(1108):20190948.
31. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging.* 2020;11(1):91.

32. Saint MJS, Orlhac F, Akl P, et al. A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study. *Magn Reson Mater Phy.* 2021;34(3):355-366.

33. Bianchini L, Santinha J, Loucao N, et al. A multicenter study on radiomic features from T2 -weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. *Magn Reson Med.* 2021;85(3):1713-1726.

34. Lee J, Steinmann A, Ding Y, et al. Radiomics feature robustness as measured using an MRI phantom. *Sci Rep.* 2021;11(1):3973.

35. Fedorov A, Vangel MG, Tempany CM, Fennessy FM. Multiparametric Magnetic Resonance Imaging of the Prostate: Repeatability of Volume and Apparent Diffusion Coefficient Quantification. *Invest Radiol.* 2017;52(9):538-546.

36. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci Rep.* 2019;9(1):9441.

37. Hoebel KV, Patel JB, Beers AL, et al. Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma. *Radiology: Artificial Intelligence.* 2021;3(1).

38. Carre A, Klausner G, Edjlali M, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Sci Rep.* 2020;10(1):12340.

39. Kickingereder P, Neuberger U, Bonekamp D, et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro Oncol.* 2018;20(6):848-857.

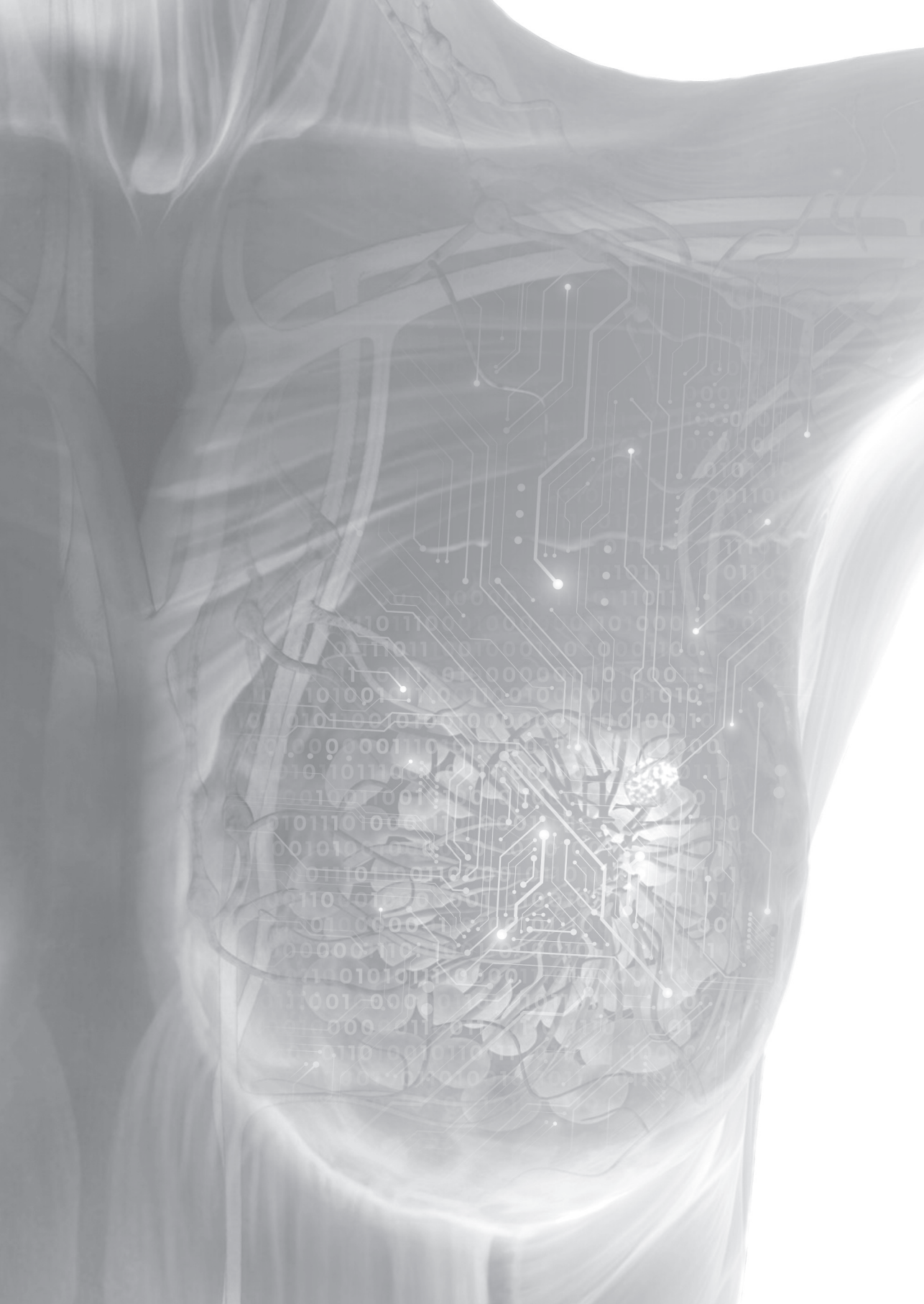
40. Shiri I, Abdollahi H, Shaysteh S, Rabi Mahdavi S. Test-Retest Reproducibility and Robustness Analysis of Recurrent Glioblastoma MRI Radiomics Texture Features. *Iranian Journal of Radiology.* 2017;Special iss(5).

41. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography.* 2016;2(4):361-365.

42. Beresford MJ, Padhani AR, Taylor NJ, et al. Inter- and intraobserver variability in the evaluation of dynamic breast cancer MRI. *J Magn Reson Imaging.* 2006;24(6):1316-1325.

43. Saha A, Grimm LJ, Harowicz M, et al. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. *Med Phys.* 2016;43(8):4558.

44. Andrearczyk V, Depeursinge A, Muller H. Neural network training for cross-protocol radiomic feature standardization in computed tomography. *J Med Imaging (Bellingham).* 2019;6(2):024008.



CHAPTER 9

Summary

This thesis consists of two parts that have the common goal of advancing personalized breast cancer care through MRI-based radiomics. Part one of this thesis investigated the use of MRI-based radiomics for prediction purposes in the treatment of breast cancer patients. Part two of this thesis focused on optimizations of MRI-based radiomics.

Part I – MRI-based radiomics for prediction purposes in breast cancer patients

In **chapter 3**, a descriptive systematic review was performed to create an overview of studies investigating the value of MRI-based radiomics for predicting tumor response to neoadjuvant systemic therapy in breast cancer patients. A total of 16 studies were included, examining data from 1,736 in total patients. The methodological quality of the included articles was judged as poor with an average radiomics quality score (RQS) of just 11% (range 0 – 41.2%). This was mainly due to the lack of external validation. In addition, radiomics methodologies showed large differences between studies, especially for tumor segmentation, feature selection, and model development, resulting in heterogeneous results that could not be compared. Nevertheless, the majority of the included articles showed promising results. Looking at the individual features, *entropy* emerged as the best performing feature with AUC values ranging from 0.83 to 0.85. The best performing multivariate prediction model scored a validation AUC value of 0.94. Based on these results, it was concluded that there is a need for standardization of the radiomics methodology to obtain comparable results in order to make further progress in this research area.

Chapter 4 investigated the possibility to predict pathological complete tumor response to neoadjuvant systemic therapy based on pretreatment MRI exams. A total of 292 breast cancer patients, with 320 breast tumors, were included in the analysis. Since the data was collected in two hospitals with five different MRI scanners and varying scanning protocols, three different strategies were used to split the data into training and validation cohorts. Radiomics, clinical, and combined models were developed and validated. The results showed that the radiomics models had no added value in predicting pathologic complete tumor response to neoadjuvant systemic therapy compared to the clinical models, nor did the combined models significantly outperform the clinical models. However, it should be noted that the effect of using data from different hospitals (with different MRI scanners and different scanning protocols on the extracted radiomic features), is still unknown. It was therefore concluded that these effects should first be investigated to determine whether further research on MRI-based radiomics for the prediction of pathologic complete response in breast cancer patients is useful.

Chapter 5 investigated whether radiomics analysis of T2-weighted dedicated axillary MRI exams can contribute to the improved diagnostic accuracy of the MRI for the prediction of axillary lymph node metastases. In this study, 511 axillary lymph nodes from 75 breast cancer patients were examined. Before the start of the radiomics analysis, all axillary lymph nodes were manually segmented in three dimensions and matched with pathology, after which 105 original radiomics features were extracted per lymph node. To validate the results, the data cohort was split into training and validation cohorts, this cohort split was performed 100 times. Each cohort split resulted in a different selection of radiomic features and with that in different AUC values. The performance of the clinical and radiomics models showed a wide range of AUC values of 0.41 – 0.74 and 0.48 – 0.89 in the training cohorts, respectively, and 0.30 – 0.98 and 0.37 – 0.99 in the validation cohorts, respectively. Based on these results, it was not possible to develop a definitive prediction model. It was concluded that radiomics analysis of dedicated T2-weighted axillary MRI exams did not contribute to the prediction of axillary lymph node metastases in breast cancer patients.

Part II – Optimization in MRI-based radiomics

Chapter 6 determined the robustness of radiomic features extracted using two commonly used radiomics software packages (RadiomiX and Pyradiomics) with respect to variability in manual breast tumor segmentations on MRI exams. A total of 129 breast tumors were segmented manually in three dimensions, by four observers: a dedicated breast radiologist, a resident, a Ph.D. candidate, and a medical student. The segmentation variability was measured using the volumetric Dice Similarity Coefficient with a mean value of 0.81 (range 0.19 – 0.96) indicating a good overlap of the breast tumor segmentations. The robustness of features was measured by the intraclass correlation coefficient. In total, 41.6% and 32.8% of all RadiomiX and Pyradiomics features, respectively, were identified as robust, independent of inter-observer manual segmentation variability.

In **chapter 7**, an MRI test-retest study was performed to assess the repeatability of radiomic features extracted from breast MRI exams. In total, eleven healthy volunteers were scanned on the same 1.5 Tesla MRI scanner in the MUMC+ using an identical scan protocol consisting of T1-weighted images, T2-weighted images, and diffusion-weighted images with corresponding ADC maps. For each healthy volunteer, 18 MRI exams were scanned on two separate days using three different test-retest strategies. From each MRI exam, the right breast was manually segmented in three dimensions where after 91 original radiomic features were extracted. Feature repeatability has been determined for features extracted from the original unprocessed images and pre-processed images. For the T1-

weighted images, the original unprocessed images were found to have the highest percentage of repeatable features, at 16.5%. The T2-weighted images showed the highest percentage of repeatable features after image pre-processing by z-score normalization, at 28.6%. The ADC maps, like the T1-weighted images, showed the highest percentage of repeatable features on the unprocessed original images, with only 8.8% of the features being repeatable. These results showed that the percentage of repeatable features in this specific setting is limited and varied per MRI sequences and image pre-processing procedure.

Samenvatting

Dit proefschrift bestaat uit twee delen met als gemeenschappelijk doel een stap verder te zetten in gepersonaliseerde borstkankerzorg met behulp van op MRI gebaseerde radiomics. Deel 1 van dit proefschrift onderzocht het gebruik van op MRI gebaseerde radiomics voor de ontwikkeling van predictiemodellen. In deel 2 van dit proefschrift werd gekeken naar de optimalisatie van MRI gebaseerde radiomics.

Deel 1 – MRI gebaseerde radiomics voor de ontwikkeling van predictiemodellen

In **hoofdstuk 3** werd een beschrijvende systematische review uitgevoerd naar de voorspellende waarde van MRI-gebaseerde radiomics voor tumor respons op neo-adjuvante systemische therapie bij borstkanker patiënten. In totaal werden 16 studies geïncludeerd waarin data van in totaal 1,736 patiënten werden onderzocht. Met een gemiddelde radiomics quality score (RQS) van slechts 11% (range 0 - 41.2%) werd de methodologische kwaliteit van de geïncludeerde artikelen als slecht beoordeeld. Dit werd met name veroorzaakt door het ontbreken van externe validatie. Er waren grote verschillen te zien in gebruikte radiomics methodologieën tussen de studies, met name in de tumor segmentatie, feature selectie en model ontwikkeling; ditwat resulteerde in heterogene, niet vergelijkbare resultaten. Desalniettemin, liet de meerderheid van de geïncludeerde artikelen veel belovende resultaten zien. Het individuele feature *entropy* liet de hoogste AUC waarde zien met een variatie van 0.83 tot 0.85. Het best presterende predictiemodel, gebruikmakend van meerdere features, had een validatie AUC waarde van 0.94. Op basis van de verkregen resultaten werd geconcludeerd dat er een noodzaak is om de radiomics methodologie verder te standaardiseren om beter vergelijkbare resultaten te verkrijgen en zo vooruitgang te kunnen boeken op dit onderzoeksgebied.

In **hoofstuk 4** werd onderzocht of het mogelijk is om door middel van radiomics te voorspellen of een tumor in de borst volledig zal verdwijnen (pathologische complete tumorrespons) na het geven van neo-adjuvante systemische therapie, op basis van MRI scans die voor aanvang van de behandeling gemaakt zijn. In dit onderzoek werden MRI scans ganalyseerd van 290 patiënten met in totaal 320 borsttumoren. Omdat de data werden verzameld uit twee ziekenhuizen waarbij er werd gescand met verschillende MRI scanners en scan protocollen, werden de te analyseren data opgesplitst in trainings- en validatie cohorten met behulp van drie verschillende strategieën. Radiomics, klinische en gecombineerde modellen werden ontwikkeld en gevalideerd. De radiomics modellen hadden geen

toegevoegde waarde in het voorspellen van pathologisch complete tumorrespons op neo-adjuvante systemische therapie op de klinische modellen. Ook presteerden de gecombineerde modellen niet significant beter dan de klinische modellen. Het effect van het gebruik van data van verschillende ziekenhuizen, verschillende MRI scanners en verschillende scanprotocollen op de geëxtraheerde radiomics features is nog steeds onbekend. Daarom werd geconcludeerd dat dit effect eerst onderzocht moest worden om te bepalen of verder onderzoek naar op MRI gebaseerde radiomics voor de voorspelling van pathologische complete respons bij borstkankerpatiënten zinvol is.

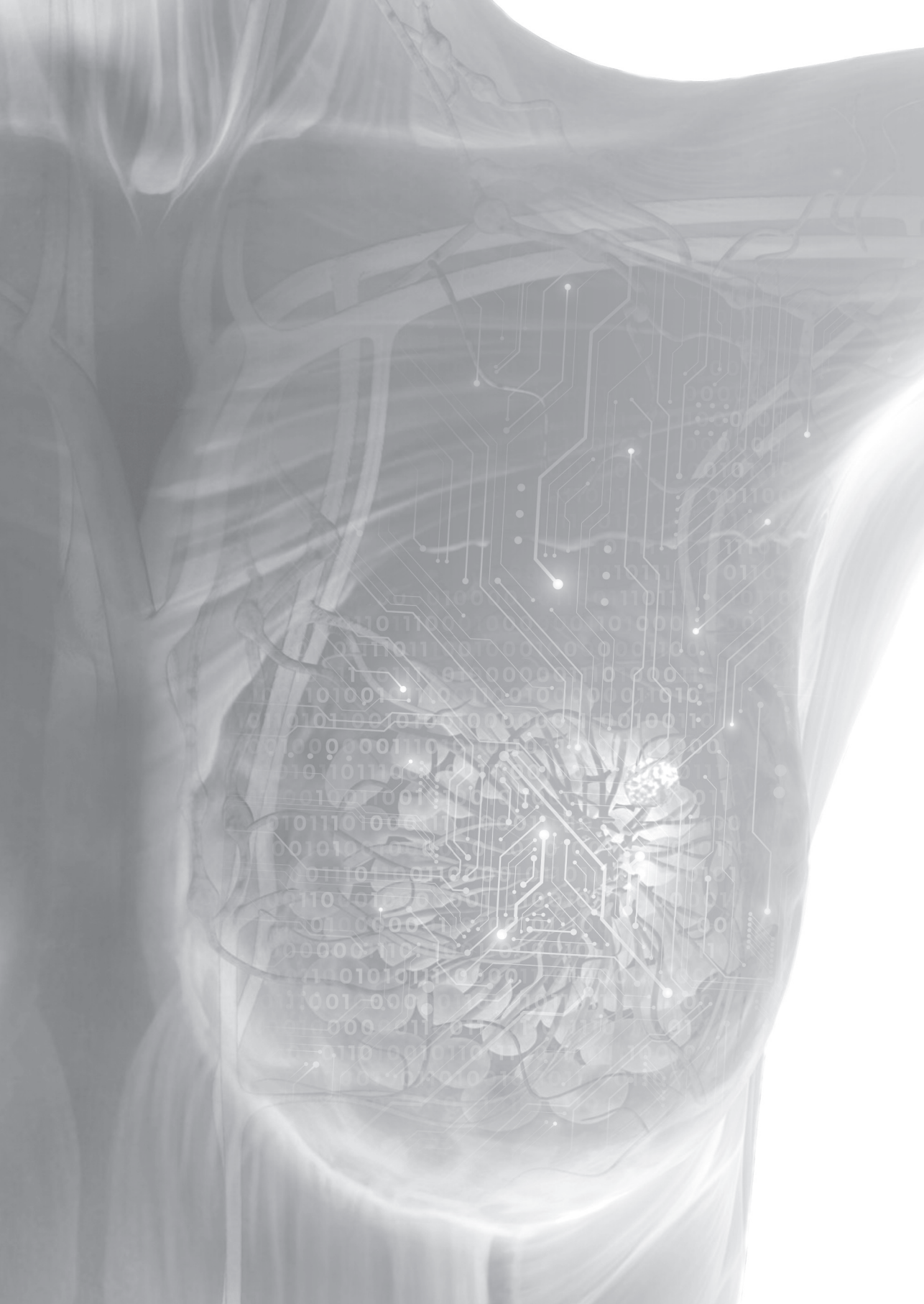
In **hoofdstuk 5** werd onderzocht of radiomics analyse van de T2-gewogen beelden van de axillaire MRI kan bijdragen aan een verbeterde diagnostische accuratesse van de MRI voor de predictie van axillaire lymfekliermetastasen. In deze studie werden 511 axillaire lymfeklieren van 75 borstkanker patiënten onderzocht. Voor de start van de radiomics analysis werden alle axillaire lymfeklieren handmatig in drie dimensies gesegmenteerd en gematcht met de pathologie, waarna per lymfeklier 105 originele radiomics features werden geëxtraheerd. Om gevonden resultaten te valideren werd het data cohort opgesplitst in een training en validatie cohort, deze cohort-splitsing werd 100 keer uitgevoerd. Elke cohort-split resulteerde in een verschillende selectie van radiomics features en daarmee ook in verschillende AUC waarden. Om de radiomics resultaten te vergelijken werden er ook klinische modellen geanalyseerd. De klinische en radiomics modellen resulteerden in een grote variatie van AUC waarden van respectievelijk 0.41 – 0.74 en 0.48-0.89 in de trainingscohorten en respectievelijk 0.30 – 0.98 en 0.37 – 0.99 in de validatie cohorten. Op basis van deze resultaten was het niet mogelijk om een definitief predictiemodel te ontwikkelen. De conclusie was dan ook dat radiomics analyse van een axillaire MRI, op basis van de gevonden resultaten, geen bijdrage levert aan de voorspelling van axillaire lymfekliermetastasen in borstkankerpatiënten.

Deel 2 – Optimalisatie van MRI gebaseerde radiomics

In **Hoofdstuk 6** werd de robuustheid van radiomics features via twee veelgebruikte radiomics software pakketten (RadiomiX en Pyradiomics) beoordeeld met betrekking tot variabiliteit in handmatig gesegmenteerde borsttumoren op MRI scans. In totaal werden er 129 borsttumoren handmatig, in drie dimensies, gesegmenteerd door vier personen: een radioloog gespecialiseerd in borsttumoren, een radioloog in opleiding, een promovendus en een geneeskundestudent. De segmentatie variabiliteit werd gemeten aan de hand van de volumetrische *Dice Similarity Coefficient*. De gemiddelde waarde hiervan was 0.81 (range 0.19 - 0.96) wat

een goede overlap van de borsttumor segmentaties indiceerde. De robuustheid van features werd gemeten aan de hand van de intraclass correlatie coëfficient. In totaal werden respectievelijk 41.6% en 32.8% van alle RadiomiX en Pyradiomics features als robuust geïndentificeerd. Deze zijn derhalve onafhankelijk van variabiliteit verkregen door handmatig uitgevoerde borsttumor segmentaties.

In **hoofdstuk 7** werd een MRI test-retest studie uitgevoerd voor de beoordeling van de herhaalbaarheid van radiomics features geëxtraheerd van MRI scans van de borsten. In totaal werden 11 gezonde vrijwilligers gescand op dezelfde 1.5 Tesla MRI scanner in het MUMC+ gebruikmakend van een identiek scanprotocol bestaande uit T1-gewogen beelden, T2-gewogen beelden, en diffusie-gewogen beelden met bijbehorende ADC-map. Van elke gezonde vrijwilliger werden 18 MRI scans gemaakt middels drie verschillende test-retest settings op twee verschillende dagen. Van iedere scan werd de rechter borst in drie dimensies gesegmenteerd, waarna per borst 91 originele radiomics features werden geëxtraheerd uit de MRI. De feature herhaalbaarheid werd getest voor zowel features geëxtraheerd uit de originele onbewerkte beelden als uit de voorbewerkte beelden. Voor de T1-gewogen beelden bleken de onbewerkte beelden het hoogste percentage herhaalbare features te hebben, namelijk 16.5%. Daarentegen vertoonden de T2-gewogen beelden het hoogste percentage herhaalbare features nadat de beelden waren voorbewerkt middels z-score normalisatie, namelijk 28.6%. De ADC-maps vertoonden, net als de T1-gewogen beelden, het hoogste percentage herhaalbare feature op de onbewerkte beelden, met slecht 8.8% van de features die herhaalbaar waren. Deze resultaten toonden aan dat het percentage herhaalbare features in deze specifieke setting beperkt is en ook varieert per MRI-sequentie en per pre-processing procedure.



APPENDICES

Impact paragraph

Main findings

This thesis investigated the use of MRI-based radiomics to advance personalized breast cancer care. First, the current knowledge was assessed through a systematic review, followed by the development of models to predict axillary lymph node metastases and pathologic complete tumor response to neoadjuvant systemic therapy. Both studies concluded that radiomics models based on MRI exams have not (yet) contributed to these predictions. However, these studies did not consider the effect of different acquisition and reconstruction parameters and the use of different MRI scanners on the extracted radiomic features, as this data was not available at the time. The estimate that these parameters affect the radiomic features is based on studies in CT imaging that showed that many radiomic features were sensitive to these effects. These studies concluded that this should be corrected before performing radiomics analysis.

In the second part of this thesis, we consequently focused on the optimization of MRI-based radiomics. One study looked at the stability of features with respect to inter-rater segmentation variability. Since automatic tumor segmentation in breast MRI is not yet sufficiently developed, studies are still dependent on manual or semi-automated tumor segmentation. For two commonly used radiomics software packages, features were identified that proved robust to manual tumor segmentation. In the MRI test-retest study, we identified a limited number of features that were repeatable regardless of the test-retest setting and scanning date for MR images used in a clinical breast protocol. These repeatable features can be used as a starting point to investigate feature reproducibility, the next step towards obtaining generalizable and comparable MRI-based radiomics results.

Relevance

Worldwide, breast cancer is the most common cancer in 2020 with 2.26 million new cases. In the same year, breast cancer ranked fifth place of most common causes of cancer death with 685,000 deaths. Breast cancer is also the most common form of cancer in the Netherlands, with each year, approximately 17,000 new breast cancer diagnoses (2). Progress in the treatment of breast cancer patients, therefore, has a major impact. Its treatment consists of surgery, systemic therapy (consisting of hormonal therapy, chemotherapy, and immunotherapy), and radiotherapy. Breast cancer is a heterogeneous disease with many variations in (non)-genetic characteristics. These variations require different treatments, ideally tailored to the individual breast cancer patient. Treatments tailored to the individual patient are called personalized medicine and have already resulted in significant progress in the treatment of breast cancer.

To advance personalized medicine, multiple sources and tools are used today, of which one is radiomics. Radiomics translates routine medical images into quantitative data that can serve as a biomarker for use in clinical decision support systems. In recent years there has been a huge increase in radiomics research, and despite mainly positive published results, incorporation of radiomics in clinical decision support systems is lagging. This is caused by various factors, including feature sensitivity to differences in acquisition and reconstruction parameters. In this thesis, the lack of this data most likely resulted in radiomics and combined models that did not contribute to the prediction of tumor response to neoadjuvant systemic therapy and radiomics and models that did not contribute to the prediction of axillary lymph node metastases.

Specifically, for breast MRI, it was even unknown whether features would remain stable when extracted from multiple scans of the same patient, scanned on the same MRI scanner with identical acquisition and reconstruction parameters. The performed MRI test-retest study gave answers to that and showed that, in the specific breast MRI setting, only a small part of the extracted features was repeatable. Most likely, further research on breast MRI-based radiomics should focus on this subset of features and examine their reproducibility. Based on this data, and after assessing feature reproducibility, radiomics analysis can be performed more reliably.

In addition, breast MRI tumor segmentation was investigated, segmentation is a necessary step before radiomics analysis can be performed. In this thesis, we identified a subset of features being robust to variability in manual tumor segmentation. As long as manual or semi-automatic segmentation is performed in breast MRI-based radiomics studies, this information can be included as a feature selection procedure.

Target population

The findings of this thesis are relevant for a broad target group; ranging from the radiomics community, technology companies and software developers, radiologists, and breast cancer patients.

- In general, reproducibility of radiomic features and the generalizability of radiomics results are issues that generate a significant amount of debate in, not limited to, the radiomics community. For that reason, many recently published radiomics articles focus on these topics and it is often concluded that reproducibility studies should be part of the data analysis itself, as it appears to be tumor site-specific. However, most of these articles use CT imaging. MR images are even more challenging since they lack the standard grayscale intensities like the Hounsfield units in CT. No test-retest study specific to breast MRI has yet been conducted, so the results of this repeatability study are a starting point for any scientist in this research field to further analyze feature reproducibility in MRI. Furthermore, this research strategy can be a source of inspiration for researchers who are investigating other tumor areas using MRI.
- The results described in this thesis may also be of interest to technology companies and their software developers, given the high reliance on software in the use of radiomics. On the one hand, for writing automatic breast MRI segmentation software, because many features were not reproducible with the still widely used manual segmentations. On the other hand, for optimizing the open-source radiomics feature extraction software, where transparency should be paramount to obtain generalizable and reproducible results.
- Although the results of this thesis are especially interesting for the scientist working in this field, it is also important that the radiologist is aware of the results of this thesis. It is ultimately the radiologist who will be using radiomics in the clinic, so it is good to involve this department in the research process early on. This also ensures that they themselves can contribute to the purpose of radiomics; assisting and supplementing the radiologist's work through clinical decision support systems.
- Ultimately, it should be the breast cancer patient who benefits from all the radiomics-related research. Although implementation in the clinic still seems a long way off, clear goals will drive progress. The application of MRI-based radiomics is likely to be first used in breast cancer diagnosis. The greatest impact on breast cancer treatment is likely to occur if the accurate prediction of tumor response becomes possible. If pathologic complete response can be predicted accurately prior to surgery, surgery and/or adjuvant therapy can potentially be omitted. In contrast, breast cancer patients who do not respond to neoadjuvant therapy can be operated on immediately.

Activities / Implementation

The results of this thesis were published in renowned international journals. Although some of them are mainly focused on the technical side of radiomics, it was decided not to publish mainly in technical journals because we think it is important that the clinician for whom radiomics will be useful in the future is already aware of the current developments. In addition, the results were presented at both national and international conferences, raising awareness among scientists working in the radiomics field, as well as clinicians working with breast cancer patients. It is especially important that radiologists, who will probably be the first users in the clinic, are involved in the development of MRI-based radiomics at an early stage so that they can also think along the implementation process. It is therefore a positive sign that the (inter)national radiological conferences are increasingly focusing on artificial intelligence, including radiomics. Furthermore, presenting our latest work in future radiomics conferences or courses like *Artificial Intelligence 4 Imaging* is an ideal way of disseminating the results to a community of leaders in the field.

Dankwoord

Aanbeland bij het laatste stukje van mijn proefschrift. Ik zeg danwel 'mijn' proefschrift, maar in de afgelopen vier jaar heb ik door hulp van velen dit mooie resultaat behaald, daarvoor mijn speciale dank in dit laatste hoofdstuk.

Marjolein, dank voor de kansen die jij mij hebt geboden. Vanuit mijn semi plek, via Lori, doorgerold in een PhD plek in jou lopende trein vol met divers mamma onderzoek. Ongelofelijk hoe makkelijk jij tijd vrij maakt in je drukke agenda om snel allerhande onderzoeksgerelateerde zaken te regelen of te bespreken, dit alles naast je drukke klinische baan en je sportieve activiteiten buiten het ziekenhuis. Super knap! Onze research journey was niet altijd makkelijk maar het was juist jou kritische blik (consistency consistency consistency ☺) en (af en toe) strenge woorden die mij enorm hebben geholpen bij maken van dit proefschrift.

Marc, ik wil je bedanken voor de fijne samenwerking. Jou kennis op het gebied van mammaradiologie in combinatie met je interesse voor artificial intelligence maakte dat ik met vragen altijd bij jou terecht kon. Jij was altijd snel en laagdrempelig te bereiken en had een kritische blik bij het reviewen van de artikelen. Tevens dank voor de vele tumorsegmentaties die je hebt gemaakt.

Henry, the first contact with the d-lab was made through you, not yet knowing that research within the d-lab would be the common thread through my thesis. You made sure that I was seen as a real D-Labber for which I am very grateful. I could always turn to your office for accessible advice. I think we can be proud of the cross pollination that has developed between the clinical mamma team and the technical d-lab team.

Leden van de beoordelingscommissie; Prof. Dr. Vivianne Tjan-Heijnen, Prof. Dr. Ruud Pijnappel, Prof Dr. Wiro Niessen, Dr. Leonard Wee. Ik wil jullie heel hartelijk bedanken voor de tijd die jullie hebben genomen om mijn proefschrift te beoordelen.

Ik wil graag alle co-auteurs hartelijk bedanken voor de tijd die eenieder heeft geinvesteerd om de artikelen te reviewen en daarmee naar een hoger niveau te tillen.

Smidties Titties, ja zo mag ik ons wel noemen!

Lieve Sanaz, oftwel Queeeeeen, wat ben ik blij dat jij en je roomies hebben gesmeekt of ik naar kamer 5.449 wilden komen ;) maar goed dat ik over stag ben gegaan. Wat hebben we gelachen, gezongen, geroddeld en uiteraard ook hard gewerkt. en niet te vergeten, ontelbare keren dubbel gelegen om de oneliner "... but that's something

else". Met onze mannelijke franse roomies maakte we het vaak erg gezellig! Sanaz, ik ben dankbaar voor alle hulp die ik van jou heb gekregen, jou tomeloze geduld en energie voor de perfecte engels zinnen zal me altijd bij blijven. Ik kon altijd rekenen op je hulp. Onze gezamelijk trip naar Athene voor het EUSOBI congress zal ik ook niet snel vergeten. Queen, ik hoop dat deze PhD een start is voor een lange vriendschap. Romy, de microbiota studies mogen blij zijn met zo iemand als jou aan het roer, altijd alles tot in de puntjes geregeld en uitgewerkt. Van deze perfectie heb ik ook mogen profiteren gezien ik (na jou 2 zwangerschappen) qua planning precies achter jou aanliep waardoor ik je vaak om advies kon vragen! Bedankt voor de altijd gezellige koffie dates, stapavonden en sleepovers! Kees, eindelijk weer een man die ons mamma-team kwam versterken, je kookskills gaven Sanaz en mij de doorslag ;) We moesten er even op wachten maar hebben er uiteindelijk van kunnen genieten, en goed dat het was!! Qua onderzoek produceerde je heel snel een aantal mooie artikelen, knap! Ik hoop dat je snel je boekje kunt afronden en ik wens je veel succes met je verdere klinische carriere. Janine, van studente rolde je zo door in je PhD traject. Jou ervaring in het lab heeft de microbiota studies een mooie boost gegeven. Dank voor de gezellige momenten die we samen hebben mogen meemaken, onder andere jou prachtige bruiloft! Lidewij, ongeveer op het zelfde moment begonnen wij ons PhD bestaan. Samen gingen we op congres in de dierentuin en gaven we onderwijs aan expats kids, olijven zoeken in kipfilets. Evie, de nieuwe chief van de BOOG studie. Dank voor de koffie momenten en gezelligheid tijdens de borstkanker evenementen. En zowaar heb ik je kunnen overhalen om ons atletisch voetbalteam te komen versterken! Lars, Sabine, Veerle en Roxanne, dank voor de gezellige dinertjes en vlaai en koffie momenten op de uni, helaas niet in grote aantallen door dat ellendige virus. Ik wens jullie allemaal heel veel succes met het vervolg van jullie PhD en hopelijk tot snel.

Thiemo, dank voor jou altijd tomeloze geduld en kritische blik op de artikelen. Jou planningskills hebben mij bij de start van mijn PhD enorm geholpen, het altijd twee stappen vooruit denken om nooit stil te vallen en de onderzoekstrein te laten rollen waren van grote waarde. Succes met je radiologische en onderzoeks carriere. Briete, dank voor de begeleiding tijdens mijn eerste paar maanden als PhD'er. Tevens wil ik je bedanken voor het eerste contact met het d-lab, dit bleek echt een springplank voor mijn onderzoek. Marissa, dank dat ik bij jou altijd terecht kon voor vragen. Heel wat leuke momenten samen mogen meemaken van een trip naar Barcelona tot etentjes in Maastricht. Thanks Maaaaris (zo zong iemand dat ooit ;)).

My French M4I roommates Fred and P-max. I'm so glad I have joined you guys, sharing the office with you was truly fun. There was always something going on when the four of us were in the office, we exchanged French and Dutch lessons, shot with nerf guns and doubled up with laughter as p-max ran down the hall

again because his beeper went off. And let's not forget our trips outside the office, I guess everyone remembers that one sushi date well. Merci pour le bon moment et j'espère te revoir bientôt!

Omdat ik bang ben om iemand te vergeten wil ik bij deze alle collega's van het heelkunde lab en M4I bedanken voor de gezellig en leerzame tijd tijdens mijn PhD. Helaas heb ik vele van jullie in mijn laatste 1.5 jaar door Corona niet meer lijfelijk gezien.

The D-lab; in het bijzonder Prof. Dr. Lambin, ik wil u heel hartelijk danken voor alle mogelijkheden die ik heb gekregen binnen het D-lab. Dit boekje was nooit tot stand gekomen zonder de hulp die ik vanuit uw kant heb mogen ontvangen. De leerzame journal clubs, de weekelijkse meetings, de 3-daagse Radiomics meeting maar ook de altijd gezellige Thembi borrels. Zelfs aan activiteiten buitenshuis als lasergamen en escape rooms mocht ik deelnemen. Also thanks to my Dlab colleagues, Relinde, Lisa, Janita, Simon, Manon, Iva, Guangyao, Fadila, Cary, Ralph, Floor, Rianne and Sebastiaan thanks for letting me be part of your team. I wish everyone the best of luck with everything to come.

Abdalla, wat ben ik blij dat jij ervoor hebt gekozen om je PhD bij het Dlab te starten. Vele uren heb ik met jou samen achter de laptop doorgebracht om mijn vele programmeer 'errors' op te lossen. Waar ik uren kon zoeken naar het programmeerfoutje had jij vaak binnen no time de oplossing voorhanden. Mijn inziens heb jij het radiomics onderzoek met speciale aandacht voor reproducibility naar een hoger niveau getild en ben ik dankbaar dat ik daar ook deel vanuit heb mogen maken. Jij hebt echt in een razendsnel tempo veel, nuttig bijdragend onderzoek verricht. Heel veel succes in NYC en ik hoop dat we elkaar nog eens tegen het lijf mogen lopen. Wellicht in NYC!

Also a special thanks to you Sergey. Your MRI preprocessing scripts really helped me through a real MRI processing battle. Everything that was easy with CT images was complicated for MRI exams, a problem we've talked about endlessly during our weekly consultations. Besides serious research talk, these meetings were always very pleasant. Sergey thanks for all your help, I wish you the best of luck with whatever comes your way.

De MRI afdeling van de radiologie in het MUMC, MRI laboranten en Roland voor de medewerking aan mijn onderzoek en het inplannen van scantijd. Speciale dank aan Liesbeth en aan Renee voor jullie meedenken en opzetten van de scanprotocollen. En dan nog dubbele dank aan Renee voor de gezellige vroege ritjes naar Maastricht en uren scantijd op de MRI. Tevens ook heel veel dank aan de proefpersonen die voor mij heel wat uren in de MRI hebben doorgebracht, zonder jullie geen onderzoek.

Het kankeronderzoeksfonds limburg, dank voor de financiele steun en gezelligheid tijdens activiteiten/borrels.

Heelkunde collega's in het Zuyderland, dank voor de warme ontvangst in jullie team en dank voor het wegwijs maken in de kliniek. Dank ook aan de Zuyderland collega's van de radiologie voor het beschikbaar stellen van de MRI beelden, met speciaal dank aan radioloog dr. Frans-Jan Hulsmans.

Dr. van Bastelaar, beste James, jouw SAM-studie was de start van mijn onderzoekscarriere. Als 'SAM-meisje' heb ik veel geleerd zowel op onderzoeksvlak als in de kliniek, waarvoor ik je hartelijk wil danken. En hoe mooi is het dat ik na mijn Maastrichtse uitstapje terug ben bij de heelkunde in het Zuyderland en jij gaat plaatsnemen in de oppositie.

Lotte, Lot, toch waal biezonder det oze paden al 30 joar lang blieve kruuse. Samen in de wieg, same noa de middelbare sjool, oeteindelijke same geneeskunde gestudeerd en tegeliekertied 'n PhD gedoan. Altied gespreksstof, tied om te lache maar auch om serieuze zake te bespreake; koffie's, lunch, high-tea time, es vanouds gezellig. En bovenal heerlijk om ff oze frustraties kwiet te kinne, wantja 'n PhD geit neet altied euver roze he. Ich hoap det oze pade nog vaak moge kruutse en det we nog veule avonture moge beleave. Heel veul succes mit dien litste loodjes!!

Thijs, Aleksandra, Anna, Kim en Bart (aka het circus). Een altijd leuke afleiding van het soms lastige PhD bestaan was het circus in Oegstgeest, welke zich inmiddels heeft verplaatst naar het prachtige Leiden. Dank voor deze welkome afleiding! ☺ En Thijs, zulle we mr zigge dat we noe qua papieren weer geliek stoan ;)

Pap en mam, waat woor ich zonger uch. Vaders, dank voor dien altied nuchtere blik op alles, t wore den waal neet de medische of onderzeuks vroage die bie dich terecht kwame (al hubse die waal motte aanheure) mr auch het leave neave mien PhD ging veuroet en doaveur hub ich dich mr al te duk veur hulp en adviezen gevroagd. Mutti, die medische vraogstukke kome bie dich terech, doa hubbe we heel get oere euver kinne vertille aan 't oavendeate of tijdens lunches en diners. Auch artikel controles of 'n engelscheck dreijde se dien handj neet veur om. Heel erg bedanktj. Maar boavenal bedankt det ger mich same altied hubtj ongersteund! Ger zeent de biste ouwers die ich mich kin winse!

De litste zeen de biste zigge ze he! Stephan, al heel get joare samen, en hoapelijk nog heel veel same te goan. Dank veur dien onuitputtelijke vertrouwe in mich tijdens de afgelaupe 4 joar. En dank detse zonger te moppere duks mien gespuij aan frustraties hubs aangeheurd. Noe is 't tied veur nuuje avonturen en doa hub ich heel veul zin in. Ich houj van dich!

List of publications

This thesis

William Rogers, Sithin Thulasi Seetha, Turkey A. G. Refaee, Relinde I. Y. Lieverse, **Renée W. Y. Granzier**, Abdalla Ibrahim, Simon A. Keek, Sebastian Sanduleanu, Sergey P. Primakov, Manon P. L. Beuque, Damiënne Marcus, Alexander M. A. van der Wiel, Fadila Zerka, Cary J. G. Oberije, Janita E van Timmeren, Henry C. Woodruff, and Philippe Lambin. Radiomics: from qualitative to quantitative imaging. BJR 2020; 93: 110820190948.

Renée W.Y. Granzier, Thiemo J.A. van Nijnatten, Henry C. Woodruff, Marjolein L. Smidt, Marc B.I. Lobbes. Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review. Eur J Radiol 2019 Dec;121:108736.

Renée W.Y. Granzier, Abdalla Ibrahim*, Sergey P. Primakov*, Sanaz Samiei, Thiemo J.A. van Nijnatten, Maaike de Boer, Esther M. Heuts Frans-Jan Hulsman, Avishek Chatterjee, Philippe Lambin, Marc B.I. Lobbes, Henry C. Woodruff, Marjolein L. Smidt. MRI-based radiomics analysis for the pretreatment prediction of pathologic complete tumor response to neoadjuvant systematic therapy in breast cancer patients: a multicenter study. Cancers 2021, 13, 2447. *Both authors contributed equally to this manuscript.

Sanaz Samiei*, **Renée W. Y. Granzier***, Abdalla Ibrahim, Sergey P. Primakov, Marc B. I. Lobbes, Regina G. H. Beets-Tan, Thiemo J. A. van Nijnatten, Sanne M. E. Engelen, Henry C. Woodruff, Marjolein L. Smidt. Dedicated Axillary MRI-Based Radiomics Analysis for the Prediction of Axillary Lymph Node Metastasis in Breast Cancer. Cancers 2021, 13(4), 757. *Both authors contributed equally to this manuscript.

Renée W. Y. Granzier, Nienke M. H. Verbakel, Abdalla Ibrahim, Janita E. van Timmeren, Thiemo J. A. van Nijnatten, Ralph T. H. Leijenaar, Marc B. I. Lobbes, Marjolein L. Smidt, Henry C. Woodruff. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. Sci Rep 2020 Aug 25;10(1):14163.

Renée W.Y. Granzier, Sergey P. Primakov*, Abdalla Ibrahim*, Sanne M. E. Engelen, Marc B.I. Lobbes, Philippe Lambin, Marjolein L. Smidt, Henry C. Woodruff. Test-Retest data for the assessment of breast MRI radiomics feature repeatability. J Magn Reson Imaging. 2021 Dec 22. *Both authors contributed equally to this manuscript.

Other

Abdalla Ibrahim, Turkey Refaee, Sergey Primakov, Bruno Barufaldi, Raymond J. Acciavatti, **Renée W. Y. Granzier**, Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Joachim E. Wildberger, Philippe Lambin, Andrew D. A. Maidment. Reply to Orlhac, F.; Buvat, I. Comment on "Ibrahim et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. Cancers 2021, 13(12), 3080.

De Rooij L, Kimman ML, van Kuijk SMJ, Granzier RWY, Hintzen KFH, Heymans C, Theunissen LLB, van Haaren ERM, Janssen A, Vissers YLJ, Beets GL, van Bastelaar J. Economic evaluation of flap fixation techniques after mastectomy: results of a double-blind randomized controlled trial (SAM-Trial). Submitted to British Journal of Surgery

Sergey P. Primakov, Abdalla Ibrahim, Janita E. van Timmeren, Guangyao W, Simon A. Keek, Manon Beuque, **Renée W. Y. Granzier**, Madeleine Scrivener, Sebastian Sanduleanu, Esma Kayan, Jianlin Wu, René Monshouwer, Hester A . Gietema, Lizza E. L. Hendriks, Olivier Mori, Arthur Jochems, Henry C. Woodruf, Philippe Lambin. Validated fully automated detection and segmentation of non-small cell lung cancer on computed tomography images. Nat Commun 13, 3423 (2022).

Abdalla Ibrahim, Turkey Refaee, Sergey Primakov, Bruno Barufaldi, Raymond J. Acciavatti, **Renée W. Y. Granzier**, Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Joachim E. Wildberger, Philippe Lambin, Andrew D. A. Maidment. Reply to Orlhac F. on "The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization". Cancers 2021, 13(12), 3080.

Abdalla Ibrahim, Turkey Refaee, Sergey Primakov, Bruno Barufaldi, Raymond J. Acciavatti, **Renée W. Y. Granzier**, Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Joachim E. Wildberger, Philippe Lambin, Andrew D. A. Maidment. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. Cancers 2021, 13(8), 1848.

Lisa de Rooij, Sander M.J. van Kuijk, **Renée W. Y. Granzier**, Kim F. H. Hintzen, Cathelijne Heymans, Lotte L. B. Theunissen, Erik M. von Meyenfeldt, Jeroen A. van Essen, Els R. M. van Haaren, Alfred Janssen, Yvonne L. J. Vissers, Geerard L. Beets, James van Bastelaar. Reducing Seroma Formation and Its Sequelae After Mastectomy by Closure of the Dead Space: A Multi-center, Double-Blind Randomized Controlled Trial (SAM-Trial). Ann Surg Oncol (2020).

Abdalla Ibrahim, Sergey P. Primakov, Manon Beuque, Henry C. Woodruff, Iva Halilaj, Guangyao Wu, Turkey Refaee, **Renée W. Y. Granzier**, Yousif Widaatalla, Roland Hustinx, Felix M. Mottaghy, Philippe Lambin. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods. 2020 Jun 3. S1046-2023(20)30111-0.

Renée W. Y. Granzier, James van Bastelaar, Sander M. J. van Kuijk, Kim F. H. Hintzen, Cathelijne Heymans, Lotte L. B. Theunissen, Els R. M. van Haaren, Alfred Janssen, Geerard L. Beets, Yvonne L. J. Vissers. Reducing seroma formation and its sequelae after mastectomy by closure of the dead space: The interim analysis of a multi-center, double-blind randomized controlled trial (SAM trial). The breast 46 (2019) 81-86.

James van Bastelaar, **Renée W.Y. Granzier**, Lori van Roozendaal, Sander M. J. van Kuijk, A. V. Lerut, Geerard L. Beets, Mo Hadfoune, Steven Olde Damink, Yvonne L. J. Vissers. Analysis of TNF-α and interleukin-6 in seroma of patients undergoing mastectomy with or without flap fixation: is there a predictive value for seroma formation and its sequelae? Surgical Oncology 28 (2019) 36-41.

Sanaz Samiei, Babbe N. van Kaathoven, Liesbeth Boersma, **Renée W. Y. Granzier**, Sabine Siesling, Sanne M. E. Engelen, Linda de Munck, Sander M. J. van Kuijk, René R. J. W. van der Hulst, Marc B. I. Lobbes, Marjolein L. Smidt, Thiemo J. A. van Nijnatten. Risk of Positive Sentinel Lymph Node After Neoadjuvant Systemic Therapy in Clinically Node-Negative Breast Cancer: Implications for Postmastectomy Radiation Therapy and Immediate Breast Reconstruction. Ann Surg Oncol. 2019 Nov;26(12):3902-3909.

James van Bastelaar, **Renée W.Y. Granzier**, Lori van Roozendaal, Geerard L. Beets, Carmen D. Dirksen, Yvonne L. J. Vissers. A multi-center, double blind randomized controlled trial evaluating flap fixation after mastectomy using sutures or tissue glue versus conventional closure: protocol for the Seroma reduction After Mastectomy (SAM) trial. BMC Cancer (2018) 18:830.

James van Bastelaar, Lori van Roozendaal, **Renée W.Y. Granzier**, Geerard L. Beets, Yvonne L. J. Vissers. A systematic review of flap fixation techniques in reducing seroma formation and its sequelae after mastectomy. Breast Cancer Research and Treatment volume 167, pages 409–416 (2018).

A

Curriculum Vitae

Renée Granzier was born on July 2th in Sittard, the Netherlands. After graduating from the secondary school Connect College in Echt, she started her bachelor's in biomedical Engineering and completed the foundation course. In 2011 she was admitted to medical school and started her medical study at the Faculty of health, medicine and life science of Maastricht University. During her bachelor's degree, she came into contact with research for the first time through a part-time job in the radiotherapy department, to which she mainly contributed by performing organ segmentations. For her last internship she was involved in a study on seroma reduction after mastectomy, under the supervision of dr. James van Bastelaar, which resulted in several publications in international peer-reviewed journals. In September 2017, directly after obtaining her medical doctor degree, she started working as a PhD student at the GROW School for Oncology and Developmental Biology under supervision of Prof. dr. Marjolein Smidt, dr. Marc Lobbes and dr. Henry Woodruff, which resulted in this thesis. Her research focused on the optimization of MRI-based radiomics in breast cancer patients for performing prediction analyses. During her PhD project she attended numerous national and international conferences and gave several talks. In September 2021 she started working as surgical resident at the Zuyderland Medical Center in Sittard and Heerlen.

A