2021•2022

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

Evidence based implementation of automated mammographic positioning quality assessment in breastcancer screening mammograms

PROMOTOR :

Prof. dr. Brigitte RENIERS

PROMOTOR :

Prof. ir. Hilde BOSMANS

BEGELEIDER :

Dr. Lesley COCKMARTIN

Marthe Picard

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleair en medisch

Gezamenlijke opleiding UHasselt en KU Leuven

2021•2022

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

Evidence based implementation of automated mammographic positioning quality assessment in breastcancer screening mammograms

PROMOTOR :

Prof. dr. Brigitte RENIERS

PROMOTOR :

Prof. ir. Hilde BOSMANS

BEGELEIDER :

Dr. Lesley COCKMARTIN

Marthe Picard

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleair en medisch

UHASSELT KU LEUVEN

Preface

This master thesis gave me the opportunity to join the medical physics in radiology team in the Medical Imaging Research Centre, MIRC, at the University Hospital Leuven. They are performing research in breast imaging in cooperation with different companies such as Volpara Health. Volpara Health has recently marketed a clinically validated, AI-powered software for breast density, which is a first step towards personalized screening and early detection of breast cancer. Next, they have developed the software Volpara TruPGMI which automatically evaluates the breast positioning quality and they are the only company to offer this type of analysis. My master thesis project focused on the performance evaluation of this software by evaluating its robustness, reproducibility and sensitivity. Furthermore, I had the exceptional opportunity to present this project at the 16th International Workshop on Breast Imaging (IWBI) congress that went on from May 22 to 25 and I am very grateful for this.

Firstly, I want to thank dr. Cockmartin Lesley and Prof. ir. Bosmans Hilde for their support, advice, and positive energy in guiding me through this thesis and for giving me the opportunity to speak at the IWBI conference. It has been a pleasure to work together with them.

Secondly, I want to thank dr. Celis Valerie, Mrs. Buelens Kristin and dr. Postema Sandra for their participation in the study and their support. Their knowledge and enthusiasm have taught me a lot and they helped me in choosing the best example mammograms for this thesis manuscript and the presentation for IWBI.

Thirdly, I want to thank Prof. dr. Reniers Brigitte for the support and revision of my work.

At last, I want to thank my family for the support and enthusiasm they showed during the whole process.

Table of contents

List of tables

List of figures

List of abbreviations

ALARA	As Low As Reasonable Achievable
CC	Cranio-caudal
CR	Computed Radiography
DR	Digital Radiography
IMF	Inframammary Fold
MLO	Medio-lateral oblique
PGMI	Perfect, Good, Moderate, Inadequate
PNL	Posterior Nipple Line
QA	Quality Assurance
QC	Quality Control
TFT	Thin Film Transistor
VDG	Volume Density Grade

Abstract

Early detection of breast cancer through mammographic screening can only be achieved with high quality mammograms. The radiologist monitors the image quality subjectively by evaluating the positioning of the breast. Volpara Health (Wellington, New Zealand) has developed a software tool called Volpara TruPGMI that does this positioning quality check automatically. Prior to its application, the reproducibility and sensitivity of the software need to be investigated.

In this study a radiologist and radiographer scored 127 screening exams with MLO and CC views of left and right breasts using 18 different positioning quality criteria. This subjective evaluation was compared to the different Volpara TruPGMI metrics. The reproducibility and compatibility of the software on the mammography systems of different vendors was verified by applying the software to mammograms of GE, Hologic and Siemens systems of different datasets. The impact of patient or technical parameters on the positioning quality was also evaluated.

The radiographer had an overall better agreement with the software then the radiologist. Additionally, the results show that the radiologist and the software did not agree on the ultimate score of 'inadequate'. The software was not able to reproduce the outcomes for different mammographic systems. Compression force and pressure had a significant influence on the positioning score. Further fine-tuning of the positioning metrics of the software towards the radiologist's decision making is part of future work.

Abstract in Dutch

Vroege opsporing van borstkanker door mammografische screening kan alleen worden bereikt met hoge kwaliteit mammogrammen. De radioloog controleert de beeldkwaliteit subjectief door de positionering van de borst te evalueren. Volpara Health (Wellington, Nieuw-Zeeland) heeft een softwaretool ontwikkeld, Volpara TruPGMI genaamd, die deze kwaliteitscontrole van de positionering automatisch uitvoert. Robuustheid, reproduceerbaarheid en gevoeligheid van de software werden onderzocht.

In deze studie scoorden een radioloog en technoloog 127 mammografische screeningonderzoeken met behulp van 18 verschillende criteria voor positioneringskwaliteit. Deze subjectieve beoordeling werd vergeleken met de objectieve beoordeling. Beelden van GE-, Hologic- en Siemens-systemen werden door de software gescoord om de reproduceerbaarheid te verifiëren. De impact van patiënt- of technische parameters op de positioneringskwaliteit werd ook geëvalueerd.

De technoloog had een algemeen goede overeenkomst met de software in vergelijking met de radioloog. Bovendien tonen de resultaten aan dat de radioloog en de software geen overeenstemming hadden betreffende de beelden met een inadequate kwaliteit. Compressiekracht en -druk hadden een significante invloed op de positioneringskwaliteit. De software is niet in staat de uitkomsten voor verschillende mammografische systemen te reproduceren. Verdere verfijning van de positioneringsmetriek van de software t.o.v. van de besluitvorming van de radioloog maakt deel uit van toekomstig werk.

1 Introduction

Breast cancer is currently the most frequent cancer in women in Europe. Through screening, systematic early detection can be achieved with the primary aim to reduce mortality from breast cancer [1]. Population screening for breast cancer has been initiated in Belgium since 2001. Women between the ages of 50 and 69 receive an invitation every two years for their mammography examination. Two images per breast are acquired, a cranio-caudal (CC) and a medio-lateral oblique (MLO) view [2]. Early detection of breast cancer through mammographic screening can only be achieved with high quality mammograms. It is the role of the radiographer to produce high quality mammograms to allow maximum visualisation of the breast tissue. It is therefore of utmost importance that radiographers in breast cancer screening programs continuously check and retrain their positioning skills. These high quality mammograms enable radiologists to detect the smallest abnormalities or subtle changes over time. In the assessment of image quality, evaluation of the positioning of the breast is a critical aspect. It is necessary to achieve uniformity and reproducibility of the screening examination [3]. In our current screening programme this positioning quality monitoring is performed subjectively by the radiologists during second reading and is based on a limited number of criteria. Volpara Health (Wellington, New Zealand) has developed a software tool for automatic and objective evaluation of the positioning quality namely Volpara TruPGMI. The software employs a scoring system, namely the PGMI standard evaluation system (which stands for perfect, good, moderate and inadequate), to assess the positional quality of each screening mammogram image. The primary focus is to evaluate whether all breast tissue is imaged. By identifying deficiencies in positioning, TruPGMI categorizes each mammographic examination (i.e., four images combined) as Perfect (P), Good (G), Moderate (M) or Inadequate (I). This results in an overall objective assessment of image quality. With a successful software tool, the radiographers would be helped with immediate feedback on the breast positioning quality and radiologists would not have to score the quality of positioning anymore. However, the experience with the software tool is still limited and we were aware of any applicable validation study. Prior to implementation into the screening practice, an independent clinical validation study is needed. This work evaluates this automatic positioning quality verification and compares it with the subjective scoring of an experienced radiologist and radiographer.

The evaluation of this software is done by examining its accuracy, reproducibility and sensitivity. By setting up a reading study with screening mammograms, a pairwise comparison to the objective software scores of positioning quality with the subjective assessment by an experienced radiologist and radiographer was made. For this subjective assessment, the criteria provided by the Dutch expert centre for screening (LRCB, The Netherlands) were used [3]. The number of images that were given identical scores were quantified as percentages of agreement and this was evaluated between radiologist, radiographer and software. In addition Cohen’s kappa values for binary data and linear weighted kappa for ordinal data were calculated for inter-rater reliability assessment. Sensitivity was examined by correctly recognizing the inadequate images.

Via the use of images of different vendors, reproducibility of the Volpara TruPGMI software could be examined. Images from other radiology practices have been collected. A comparison was made between GE, Hologic and Siemens images. The Pearson's chi square test was used to indicate significant differences. A post hoc Fisher exact test was then performed to confirm precisely where the differences were situated.

The anatomical characteristics of the breasts have also an impact on the quality of positioning. It was investigated whether parameters such as compression force, compression pressure, breast volume, Volpara density grade (VDG) and dose have an influence on this. By means of an ordinal regression it can be determined to what extent such parameters influence the overall quality of the image.

In the first chapter, an introduction about screening policies in Belgium will be given. Since Belgium follows the European guidelines for breast cancer screening and diagnosis, this will also cover the European viewpoint on breast cancer screening. Chapter two describes the main evaluation parameters for the population screening programme. Chapter three provides a brief explanation of the mammographic device. Chapter four discusses the software under investigation namely Volpara TruPGMI. Next, chapter five explains the various statistical tests that were used in this work. First, the Cohen's kappa test that determines interrater reliability is described. The Chi-square test as well as the fisher's exact test are then described which can determine whether or not two or more classifications of samples are independent. Finally, ordinal regression is described. Chapter six describes the research objectives, followed by the section on materials and method with the characterization of the different patient datasets. After that, the statistical tests are further explained as well as the setup of the reading study. Chapter seven shows the actual research results, starting with the patient demographic data, and then the comparative analysis of positioning quality assessment between software and readers. The results for the reproducibility testing using different vendors and the influence of patient specific and acquisition parameters on the positioning quality are also shown. Chapter eight discusses the possible differences between radiographer and radiologist vs software and compares them with literature findings. At last, chapter nine contains the conclusion and future outlook of this research

2 Breast cancer screening

Population screening is the examination of a basically healthy population to detect asymptomatic cases of a disease or condition, on the assumption that this condition may be better treated at an early stage [4]. The objective is to detect any abnormality in asymptomatic cases with the effect that less severe treatments need to be given and chance of recovery is greater. The primary aim is to lower mortality from breast cancer. Because breast cancer can be detected at an early stage, this disease is eligible for screening [1].

In 1988, the first edition of the document 'European Guidelines for Quality Assurance in Mammography Screening' was published. This document was part of a series of projects to fight against cancer for the European citizens, on behalf of the European Union. The guidelines address topics such as training, multidisciplinary teamwork, minimizing adverse effects, supervision and monitoring as well as cost-effectiveness. The third edition, released in 2001, became the standard for breast cancer screening in Europe. The aim is to provide a more uniform level of service so that any advances in technical and professional knowledge can be shared. In the fourth and so far last edition, published in 2006, new chapters on communication and physio-technical aspects of digital mammography have been added. Currently, at least 22 countries participate in regional and national population-based breast cancer screening programs [1]. According to the WHO and Perry et al. a 20% reduction in mortality from breast cancer could be accomplished by systematic screening. The cumulative benefit on total specific mortality may be visible as early as four years with a maximum effect twelve years after implementation of a screening program [1, 5].

Mammography remains the screening tool of choice for breast cancer screening. A mammogram is an x-ray image of the breast [6]. Early detection of breast cancer through mammographic screening can only be achieved with high quality mammograms so that sufficient diagnostic information can be provided. And this by using a radiation as low as is reasonably achievable (ALARA). Quality control (QC) is necessary to ensure that each mammogram is of high quality and is done through controls of the equipment. The equipment, consisting of modern dedicated X-ray machines and appropriate image receptors must meet accepted performance standards. Quality control of the technical and physical aspects begins at the specification and purchase of the material and before it can be implemented in the clinic it must undergo acceptance testing. This is done for the mammography X-ray equipment, the image receiver, the viewing device and the QC test equipment to ensure that the performance meets these standards. Whereas QC holds for the equipment, quality assurance (QA) in the screening program takes into account the medical, organizational and technical aspects. Most of the testing for QC can be performed by the staff but the more extensive and specific measurements must be performed by certified medical physicists. Radiographers have also a prominent role in QC as they should carry out daily phantom acquisitions. Besides QC, the radiographer can achieve high quality mammograms by correct positioning of the breast. This is necessary to allow maximum visualisation of the breast tissue, reduce recalls for technical inadequacies and maximise the cancer detection rate [1].

The radiologist reads and interprets the mammograms and must be able to detect the smallest abnormalities and subtle changes over time. They have the end responsibility that the mammograms are of high quality. They should be aware of the periodic training of the personnel and the risks and benefits of breast cancer screening. Together with the radiologist and radiographer, a multidisciplinary team is responsible for breast cancer screening, diagnosis and treatment including pathologists, surgeons and nurses, with additional input from oncologists, physicists and epidemiologists [1].

2.1 Breast cancer screening in Belgium

In Belgium, breast cancer is the most common cancer in women. In 2019 a total of 10,962 women were diagnosed with breast cancer. Breast cancer has a 5-year relative survival rate of 91.9% and it affects mostly women above the age of 50 years [7, 8]. Due to the high incidence, the Flemish government started a Breast Cancer Screening Program for women aged 50 to 69 years in 2001 organized by '*Het Centrum voor Kankeropsporing'* (CvKO). It encourages women of this age group to undergo a screening mammogram every 2 years. Two images per breast are acquired, a cranio-caudal (CC) and a medio-lateral oblique (MLO) view (figure 1). Each image is viewed by at least two radiologists, independently. In case the assessment of the two radiologists differs, a third reading is done [2].

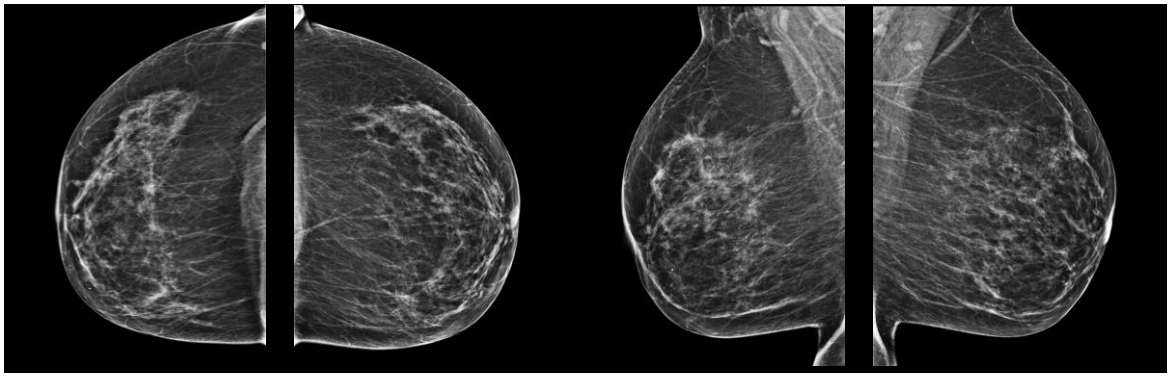

Figure 1: CC view (left image) and MLO view (right image) of left and right breast

Because Belgium consists of three regions, Flanders, Wallonia and Brussels, each has its own policy on implementation and coordination of population screening, but all are based on the European guidelines. In Wallonia it is the *'Centre Communautaire de Référence pour le dépistage des cancers'* (CCR) and for Brussels it is the *Bruprev* (formerly *Brumammo*) who are responsible for the organization of the breast cancer screening programme. For the three regions, the same stringent quality criteria are maintained for the radiological services involved. In Flanders there are 160 approved mammographic units, in Wallonia 77 and in Brussels 34. In addition to screening for breast cancer, screening for cervical and colon cancer is also done in Flanders. In Wallonia and Brussels, there is also a population screening for colon cancer but there is not yet an organized population screening for cervical cancer [9].

To be allowed to start a population study, criteria have been drawn up by the Flemish government that relate to the disease or disorder, the target group, the screening instrument, the diagnosis and treatment as well as the population study as a whole. The screening instrument is the medium, such as a survey, test or measurement, used for screening and should be simple, safe, accurate, effective and of high quality. Detectable risk factors and disease precursors must be known at an early stage, and treatment at an early stage must have a better outcome than at a late stage. The benefit to the target population should be larger than the harm caused by the population screening. All persons in the chosen target group should have the opportunity to participate in the screening and be correctly informed about it. Good communication with professionals and the target group about the course of the study, advantages and disadvantages, and the available choices is very important [10].

In addition to the obvious advantages, there are also some disadvantages to screening. Because the images are reviewed by 2 radiologists, to reduce the chances of a missed injury or interval cancer, it

can be a long wait for mammogram results. This can be unpleasant and create needless anxiety in the women. Patients in whom cancer is discovered to be growing very slowly and show clearly benign features do not always need to receive treatment, this is called overdiagnosis. This leads to non-necessary treatments and creates unnecessary anxiety in the patient [1, 2]. The radiation that the patient receives during a mammography examination is very small, namely less than 0.7 mSv. To put this in perspective, in 2015 the estimated exposure to ionizing radiation for the average Belgian by cosmic and terrestrial radiation was 0.7 mSv per year. Despite the low dose, stochastic effects can still occur at this dose but it is far below the threshold for deterministic effects [11].

2.2 Screening results and statistics

Figure 2 shows the evolution of breast cancer mortality in Flanders for all ages and for the age group 50-69 years from the year 2004 to 2015 for age-standardized incidence using the world standard population (N/100,000 person-years). There is a clear downward trend in mortality, especially for women aged 50-69 years [12].

To verify the effectiveness of a screening program, the analysis of the effect on mortality alone would require a long follow-up time. Surrogate performance indicators have been introduced such as interval cancer rate, sensitivity, cancer detection rate. These can be monitored and evaluated periodically [13]. Interval cancers, are cancers that manifest between two screening rounds after a negative previous screening examination. This leads to false reassurance in the patients who received a negative result [1, 5].

The Centre for Cancer Detection publishes annually a report with detailed numbers and statistics on screening in Flanders. The results of the year 2019 are representative figures of the screening program from before the covid period. In 2019, 397,831 invitations were sent with a response rate of 54.1%. The coverage rate was 65.1%, i.e. full screening both inside and outside the organized population study [14]. This is a much higher percentage than in the beginning of the screening program (46.2% in 2003) [13]. Of all the women in the target group, 14.5% had never participated in any kind of breast cancer screening before, which is 127,018 women. This figure appears to be very stable over the years. Partly because of this, the goal for overall coverage of 70% is not being met [14]. The participation rate has a major impact on the success rate of a population study [5]. The recall rate after a screening mammogram should be below 7% and 5% for initial screening and follow-up screening, respectively, according to European standards. The 2019 results are within these standards. A recall means that the woman must physically return to the screening centre for a new examination. The reason may be for further examination after seeing an abnormality observed in the screening examination or because of a technical defect of the screening mammogram. A technical defect may mean improper breast placement, processing errors or machine or operator errors. The difference with a repeat screening test is that the woman receives additional testing during the screening examination and does not have to be physically recalled. On-site treatment of a woman before discharge can significantly reduce, but never completely eliminate, technical recalls. Regardless of whether an examination must be repeated at a later time due to unsuitable quality for diagnostic purposes assessed by the radiologist or at the time of the screening examination due to a technical problem identified by the radiographer, all repeat examinations should be recorded and kept to a minimum of less than 1 per 100 women examined [1].

The annual report also indicates that screen-detected cancers are more often diagnosed at an earlier stage than in non-participants and interval cancers. More than half of non-participants have a more advanced stage at diagnosis. The breast cancer detection rate at initial screening and follow-up screening meets the desired European standard of ≥6,9/1.000 and ≥3,45/1.000 respectively [14].

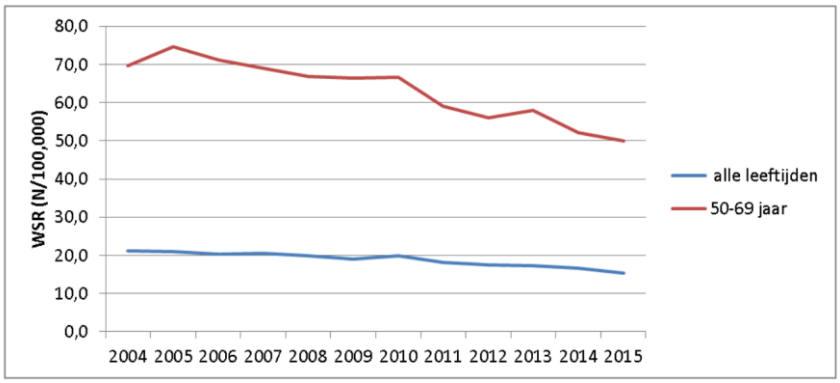

Figure 2: Evolution of breast cancer mortality in Flanders for all ages and for the age group 50-69 years (2004-2015)[1] [12]

[1] WSR: age-standardized incidence using world standard population (N/100,000 person-years)

3 Mammography

A mammogram is an x-ray image of the breast to detect breast cancer or other breast diseases and can be used both as a diagnostic and screening tool. A mammography system consists of an X-ray tube, filter, collimator, compression plate, scatter grid and finally the detector (figure 3).

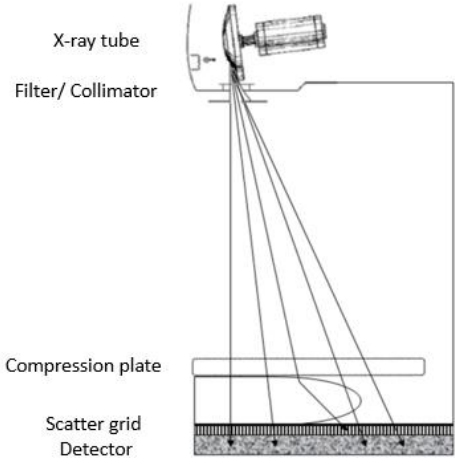

Figure 3: Schematic representations and components of a mammographic system [15]

3.1 X-ray production in an x-ray tube

The X-ray tubes used in mammography devices use the same principles as in conventional radiography. One works with an anode and cathode to generate radiation. The cathode is made of tungsten and by putting it under current, free electrons are generated. The negative free electrons are attracted and accelerated by the anode which is at a positive higher voltage than the cathode. The anode usually consists of molybdenum or tungsten. The voltage difference between the anode and cathode is called the tube voltage and is typically 28 kV in mammography. The electrons collide in the anode material and create bremsstrahlung and characteristic radiation. Bremsstrahlung is the slowing down of the electrons in the anode material which creates photons (figure 4). Not every electron is slowed down equally and so each photon gets a different energy, a spectrum of photons with different energies is thus created. Characteristic radiation occurs after the emitted electron collides with another electron from an atom. This then creates a vacancy in the shell of the atom. This vacancy is filled by a higher lying electron and the radiation released in the process always has the same energy. This is because the electron orbitals of an atom are always at the same energy level (figure 5) [15].

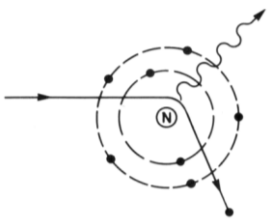

Figure 4: Bremsstrahlung [15]

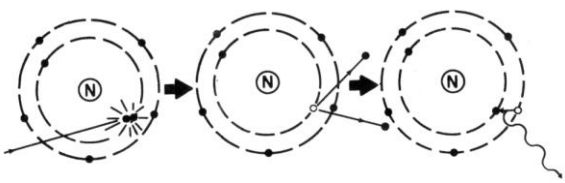

Figure 5: Characteristic radiation [15]

3.2 Filter, collimator and compression

A filter aims to reduce the dose to the patient without loss or limited loss of contrast. In digital mammography, this filter consists most often of silver or rhodium which are used for their K-edge attenuation. In this process, the absorption above the K-edge increases which attenuates the hard components (high energy radiation) from the x-ray beam. The collimator delimits the X-ray beam and ensures that the radiation beam does not fall outside the bucky and detector. Hence, X-rays that fall outside the bucky on the thoracic side provide a higher dose to the patient without contributing to imaging and this should be maximally avoided [15].

During a mammographic exam, the breast is compressed between the flat support plate and a paddle which is a plate parallel to the support plate. Because the compressed breast is thinner than a non-compressed breast, a shorter absorption time is required and thus less radiation. Less scatter radiation will also be generated which all improve the image quality [15]. Keeping the breast in a fixed position will also reduce the risk of motion artifacts and image blur. One drawback is that it can be uncomfortable and painful [16].

3.3 Digital detectors

Many different DR-systems are nowadays available and in development. A flat plate CsI with photodiode array is a common used system. In this system, a CsI(Tl) phosphor layer is used that converts the X-rays into light. This phosphor layer is deposited on a matrix of photodiodes consisting of amorphous silicon. The light photons emerging from the scintillator are converted to charge in the photodiodes via a thin film transistor (TFT) switch. Digitizing this creates the image. Each photodiode corresponds to a certain pixel value [6, 15].

A second system is the flat plate amorphous selenium with electrode array. This system does not use phosphor. Instead, a selenium layer is created where, through interaction with X-rays, electron-hole pairs are formed. By means of an applied electric field over the selenium layer, these electron-hole pairs can be collected and transported in the electrodes. The reading of this is done by TFT switches placed on each detector element [6].

4 Volpara software

Volpara TruPGMI (Volpara Health, New Zealand) is a commercially available software to assess the positioning quality of every screening mammographic image. The primary focus is to evaluate if all breast tissue is imaged. To be able to give a score to an image, the software first determines and segments landmarks and measures distances or angles of interest. The measurements are then gathered into TruPGMI Metrics where different weightings are given to the metrics. These metrics are in turn collected into an Image Score. This score is the PGMI score and stands for Perfect (P), Good (G), Moderate (M) or Inadequate (I). Via weighting of the Image Score of the 4 images that are part of the mammographic exam, a total Study PGMI score is obtained, which is again a PGMI score. The final results is an overall objective assessment of image quality, positioning wise. A moderate image is of acceptable quality but some improvements need to be considered in terms of positioning whereas an inadequate image is of poor quality. These scores provide immediate feedback to the radiographer on his/her positioning performance. Images with poor scores might require a retake after consideration or on request by the radiologist, who can interpret the need for an improved positioning and/or the feasibility to obtain a better image in a specific patient. Four metrics for CC view and seven metrics for MLO view are determined by Volpara TruPGMI (table 1).

Table 1: TruPGMI metrics for CC and MLO view

TruPGMI metric		
CC	1	Nipple in profile
	2	Nipple in midline of imaged breast
	3	Posterior nipple line (PNL) within 1 cm of PNL on MLO view
	4	No cut off of breast tissue
MLO	1	Nipple in profile
	2	Inframammary fold (IMF) visible
	3	No cut off of breast tissue
	4	Pectoral inferior muscle length
	5	Pectoral muscle shape
	6	Pectoral muscle adequacy
	7	No pectoral skin folds

Figure 5 indicates the various metrics that will be discussed. The first criterion for CC view is about the nipple in profile. The nipple is for both views, CC and MLO, the anterior reference point and should protrude the skin line. The second criterion includes also the nipple. Besides being in profile it should lay in de midline of the imaged breast. When the nipple deviates from the midline, the software gives an output according to the size of the deviation. When the deviation is small, the output is "exaggerated" and when the deviation is large, "too excessive". The third criteria is about the posterior nipple line (PNL). The PNL is the distance from the nipple to the posterior side of the breast and is a way of measuring whether enough tissue is visible. The PNL on CC view should be within 1 cm of the PNL on MLO view. If this criterion is not met then it can be inferred that the CC image does not show enough breast tissue. The last criterion evaluates whether there is tissue cut off or not. This occurs when the breast is not centrally placed on the bucky so that the medial or lateral part of the breast is not visualized.

In MLO view, the nipple is also examined for its profile. Cut off in MLO view occurs only at the bottom of the breast. Another criterion is the inframammary fold (IMF) that should be visible. The IMF is the

transition of the breast to the rest of the chest. When the IMF is not visible it can be obscured by a fold or by superposition of the abdomen on the breast. In the worst case the IMF is absent. The correct depiction of the pectoral muscle is the most important criterion in the MLO view. The pectoral inferior muscle length is the vertical distance between where the pectoral muscle intersects with the posterior image edge and with the anterior margin of the pectoral muscle. Together with the PNL it can be determine whether enough tissue is depicted or not. The intersection of the pectoral inferior muscle length with the posterior image edge should be within 1 cm of the intersection of the PNL with the anterior margin of the pectoral muscle or with the posterior image edge if the pectoral muscle is not depicted deep enough. The fifth criterion is about the pectoral shape, which can be convex or concave whereas convex or flat is preferred. Pectoral adequacy is determined by the depth and width of the depiction of the muscle. The depth, measured on the posterior side of the image, should be at least one-third of the total height of the chest. If the depth is shorter, this is not adequate. The width must be within the thresholds. If the muscle is not wide enough the output is "narrow" or conversely if it is too wide it will say "wide". The last criterion is whether any fold are visible in de pectoral muscle. The software is not yet able to recognize folds in breast tissue.

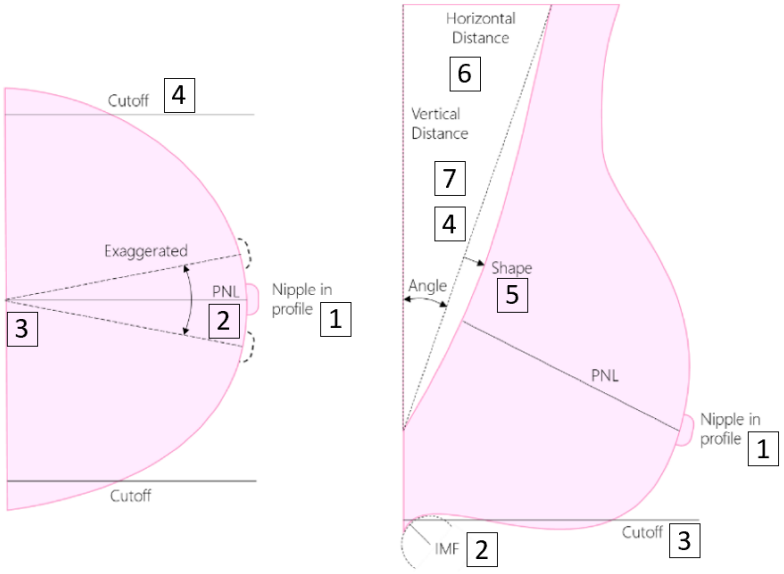

Figure 5: Volpara PGMI metrics for CC (left) and MLO view (right)

5 Materials and Method

5.1 Setup of reading study

As a comparison to the objective software scores of positioning quality, the screening mammograms were also rated by an experienced radiologist and radiographer. For this subjective assessment, the criteria provided by the Dutch expert centre for screening (LRCB, The Netherlands) were used [3]. In total 18 quality criteria for CC and MLO view were scored for each mammographic screening exam (table 2). The rating scale of some criteria was adapted in order to enable a direct comparison with the metrics provided by Volpara TruPGMI. The metrics around skin fold in CC view, depiction of the pectoral muscle in CC view and symmetry in both views could not be compared to the outcomes of the software as no metrics exists to do these measurements in the software.

Table 2: Breast positioning criteria and associated answers for the reader study and for Volpara software

	Question		*Possible answers*	*Volpara metric*	
CC view	**1**	**Folds present**	*Yes, no*	**NA**	
	2	**Depiction of pectoral muscle**	*PGMI*	**NA**	
	3	**Sufficient breast tissuelateral side**	*Yes, no*	**Tissue Cut Off**	*False, True*
	4	**Sufficient breast tissue medial side**	*Yes ,no*	**Tissue Cut Off**	*False, True*
	5	**Symmetry**	*Yes, no*	**NA**	
	6	**Nipple in profile**	*Yes, no*	**Nipple in Profile**	*InProfile, NotInProfile*
	7	**Overall quality**	*PGMI*	**Image PGMI Score**	*PGMI*
	8	**Retake needed**	*Yes, no*	**Image PGMI Score**	*PGMI*
MLO view	**9**	**Folds present**	*Yes, no*	**Pec Skinfold**	*Present, Absent*
	9.1	**If yes, where**	*Pectoral, breast*		
	10	**Depiction of inframammaryangle**	*Visible, obscured, absent*	**IMF adequacy**	*VisibleWithoutSkinfold, VisibleButWithSkinfold, NotVisible*
	11	**Pectoral muscle sufficiently deep**	*Yes, no*	**Pectoral Adequacy**	*Adequate, NotAdequate (Short)*
	12	**Pectoral muscle sufficiently wide**	*Yes, no*	**Pectoral Adequacy**	*Adequate, NotAdequate (Narrow, Wide)*
	13	**Symmetry**	*Yes, no*	**NA**	
	14	**Nipple in profile**	*Yes, no*	**Nipple in Profile**	*InProfile, NotInProfile*
	15	**Lower side of breast cut-off**	*Yes, no*	**Tissue Cut Off**	*False, True*
	16	**Overall quality**	*PGMI*	**Image PGMI score**	*PGMI*
	17	**Retake needed**	*Yes, no*	**Image PGMI score**	*I versus PGM*
	18	**Is there a quality problem with an image (i.e. compression, saturation, motion, noise,.. please indicate in which image and describe the problem in notes panel**	*Free answer*		

For the reader study, the software ViewDEX was used. ViewDEX is a Java-based software for presentation and evaluation of medical images in observer performance studies [17, 18]. The images were displayed on a 5 megapixel medical grade monitors (Barco, Kortrijk, Belgium) calibrated to the DICOM grayscale standard display function (figure 6). Each reading session contained around 25 cases, each consisting of a MLO and CC view from left and right breast. Neither prior images nor medical history were available to the readers.

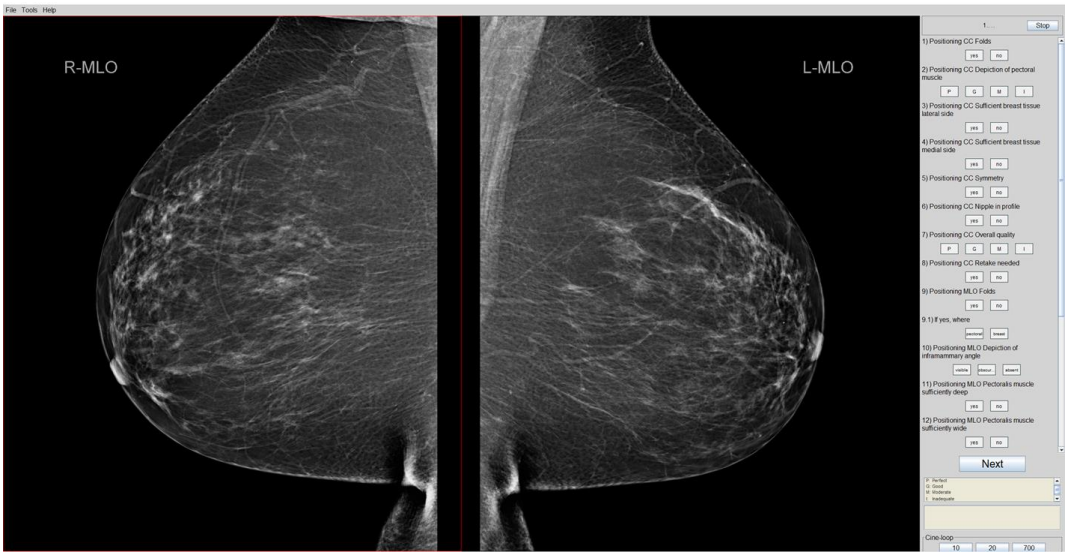

Figure 6: Screenshot of ViewDex

5.2 Patient dataset

Following IRB approval and GDPR compliance, 127 mammographic screening exams were retrospectively and randomly collected for the period of Jan 2017 and March 2021 at University Hospitals Leuven (Belgium). All examinations were performed on a Siemens MAMMOMAT Inspiration or a Siemens MAMMOMAT Revelation mammography system (Siemens Healthineers, Forchheim Germany). Volumetric breast density was assessed automatically using Volpara Density (Volpara Health, New Zealand) categorizing the images into density classes following the 5th edition of the Breast Imaging Reporting and Data System (BI-RADS).

This dataset of 127 mammographic examinations is further compared with 2 other datasets obtained from other radiology practices and therefore different technologists doing the positioning. This therefore implies other vendors too. A comparison is made between Siemens, GE and Hologic. It can be assumed that the datasets are equivalent in quality. Images from Fuji-systems could not be included as the software cannot work with detector images with a logarithmic response function of pixel intensity as a function of dose at the detector. In table 3 below, the number of examinations within the datasets for each vendor is shown.

Table 3: Total number of examination per brand

Brand	Total number of examination
GE	186
Hologic	213
Siemens	127

The influence of parameters such as breast volume, compression force and pressure, breast volumetric and mean glandular dose on the quality of breast positioning is also examined. The dataset used for this purpose was much larger. A total of 4888 mammographic images, all from a Siemens mammography system, have been collected which equates to approximately 1222 examinations or patients.

5.3 Statistical analysis

This section discusses the statistics used to process the results. More detailed explanations of the statistical tests used can be found in the annex.

Paired wise comparisons were made between the reading study output and the software output. Absolute number of images that were given identical scores were quantified as percentages of agreement and vice versa for disagreement between radiologist, radiographer and software. In addition Cohen's kappa values for binary data and linear weighted kappa for ordinal data were calculated for inter-rater reliability assessment using SPSS version 28 (IBM, Chicago, US). Correlation was interpreted according to the following distribution: <0, no agreement; 0 – 0.20, slight agreement; 0.21 – 0.40, fair agreement; 0.41 – 0.60, moderate agreement; 0.61 – 0.80, substantial agreement; 0.81 – 1.0, perfect agreement [19].

To determine if there is a significant association between PGMI distributions of the images of different vendors, a Pearson's Chi square test was performed. Two sided p-values of <.05 were considered statistically significant. The criteria about tissue cut off, nipple in profile, adequate depiction of the pectoral muscle and folds in MLO view, inframammary angle and finally the overall quality were compared. If it was found that there is a significant difference, a post hoc test namely the Fisher exact test was performed to determine where this difference lies. Since we subjected our data to a second statistical test, Bonferroni correction for multiple comparison needed to be applied. A comparison between GE and Siemens, GE and Hologic and between Siemens and Hologic was made and therefore the data were subjected to three tests which reduces the significance threshold level to .017 (p/n = .05/3).

To evaluate the possible influencing patient-specific or acquisition factors on the positioning quality, an ordinal regression analysis is carried out (SPSS version 28, IBM, Chicago, US). Patient-specific factors are the breast volume (cm^3) and Volpara density grade (VDG). The VDG is divided into four classes with increasing breast density from A to D. Acquisition factors includes compression force (N) and -pressure (kPa) and the dose (mGy). Compression pressure is the compression force applied by the paddle, divided by the area of the breast [16].

6 Results

6.1 Patient demographic data

All women were between 49 and 69 years old with an average age of 62 years. Half of the women received a BI-RADS density category B, whereas 16%, 21% and 11% of the women were in density category A, C and D resp. The compression force ranged between 70 and 188 N with an average of 118 N which is within the recommended range according to our local standards. In literature recommendations for compression pressure around 10 kPa are reported [16]. Compression pressures in this study ranged from 4 to 32 kPa with an average of 12 kPa.

6.2 Comparative analysis of positioning quality assessment between software and readers

6.2.1 Skin folds in the CC image

A direct comparison between the quality criteria scored by the radiographer (R1) and radiologist (R2) was made. Table 4 shows the results of the reader scores. The radiographer detected in 36% of the cases a skin fold in the left or right CC image whereas the radiologist detected it in 26% of all cases. The agreement between the radiologist and radiographer was 77% and the kappa value was 0.47 which corresponds with a moderate agreement. Figure 7 shows an example of a fold present in CC view and was provided by dr. Celis because no good examples were available in the dataset. Therefore it is not possible to provide the assessment of the readers on this example.

Table 4: Percentage (%) and proportion (n = out of 127 examinations) of images with skin folds in CC view detected by the radiographer and radiologist and corresponding agreement between the reader

	Yes	No
	% (n=)	% (n=)
Radiographer	36% (46)	64% (81)
Radiologist	26% (33)	74% (94)
	Agreement	Disagreement
	77%	23%
	$\kappa = 0.47$	

Figure 7: Fold present in right CC view indicated by an arrow

6.2.2 Depiction of the pectoral muscle in CC images

The pectoral muscle was rated as 'not present' in 51% of the cases by the radiographer and in 32% by the radiologist and were consequently scored as Inadequate for this metric. Furthermore the radiographer scored 13% of the cases as Moderate, 12% as Good and 24% as Perfect. This means that most cases were scored as Perfect or Inadequate, the extremes. The radiologist scored 43% as Moderate, 20% as Good and only 6% as Perfect (table 5). It can be seen that the radiologist scored most of the cases as Moderate or Inadequate. The absolute agreement between the radiologist and radiographer was therefore low, i.e. 47%, however moderate agreement was obtained with kappa statistics (κ_w =0.50). Figure 8 shows a perfect depiction of the pectoral muscle in right and left CC view, indicated by arrows. Both the radiographer and the radiologist indicated that the visibility of the pectoral muscle is good.

Table 5: Percentage (%) and proportion (n = out of 127 examinations) of pectoral depiction in CC view by the radiographer and radiologist and corresponding agreement between the readers

	P	G	M	I
	% (n=)	% (n=)	% (n=)	% (n=)
Radiographer	24% (31)	12% (15)	13% (16)	51% (65)
Radiologist	6% (7)	20% (25)	43% (55)	32% (40)
	Agreement	Disagreement		
	47%	53%		
	κ_W = 0.50			

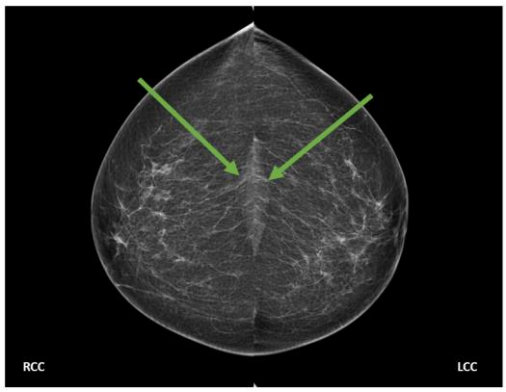

Figure 8: Depiction of the pectoral muscle

6.2.3 Left and right symmetry in CC and MLO view

Symmetry between left and right breast in CC and MLO view was also evaluated in the reader study. The percentages of images concluded as symmetrical was 80% for CC view both for radiographer and radiologist. For MLO view the radiographer and radiologist scored 83% and 87% as symmetrical respectively (table 6). The agreement between the radiologist and radiographer for CC and MLO view was 84% and 90% respectively with kappa values of 0.50 and 0.61 which corresponds with moderate and substantial agreement.

Table 6: Percentage (%) and proportion (n = out of 127 examinations) of left-right symmetry in CC and MLO view by the radiographer and radiologist and corresponding agreement between the readers

	CC view		MLO view	
	Yes	No	Yes	No
	% (n=)	% (n=)	% (n=)	% (n=)
Radiographer	80% (102)	20% (25)	83% (105)	17% (22)
Radiologist	80% (102)	20% (25)	87% (110)	13% (17)
	Agreement	Disagreement	Agreement	Disagreement
	84%	16%	90%	10%
	$\kappa = 0.50$		$\kappa = 0.61$	

6.2.4 Nipple in profile

The percentage of images with “nipple in profile” in CC view was 79%, 93% to 69% for radiographer, radiologist and software resp. In MLO view 81%, 95% and 79% was scored as “nipple in profile” by the radiographer, radiologist and software (table 7). In this study, an agreement between the radiographer and the software of 81% for CC view and 87% for MLO view was found. For the radiologist and the software an agreement of 73% for CC view and 82% for MLO view was obtained. The software seems to be more strict than the readers for this criterion. An inter-rater reliability between the radiographer and radiologist versus the software of 0.51 and 0.15 resp. for CC view, and 0.58 and 0.18 resp. for MLO view was obtained. The kappa values between the readers is 0.25 and 0.35 for CC and MLO resp. It is important that the cases scored as "not in profile" by the software are the same images to which the readers declared that the nipple is not tangential. From the cases that were scored “not in profile” by the radiologist, there were 67% in agreement with the Volpara software for both CC and MLO views. From the cases that were scored “not in profile” by the radiographer in CC view, there were 78% in agreement with the Volpara software and 71% in MLO view. In figure 9 and 10 the skin is indicated by an orange dotted line. Figure 9 is a good example where the nipples are not in profile, as the nipples lie within the skin edge. In figure 10 the nipples do protrudes the skin line and are both in profile. Radiographer, radiologist and software all agreed on the two examples.

Table 7: Percentage (%) and proportion (n = out of 127 examinations) of nipple in profile for CC and MLO view by the radiographer (R1), radiologist (R2) and Software (S), and the agreement, disagreement between radiographer and Volpara (R1- S) and radiologist and Volpara (R2-S)

	Nipple in profile – CC view		Nipple in profile – MLO view	
	Yes	No	Yes	No
	% (n=)	% (n=)	% (n=)	% (n=)
R1	79% (100)	21% (27)	81% (103)	19% (24)
R2	93% (118)	7.1% (9)	95% (121)	5% (6)
S	69% (88)	31% (39)	79% (100)	21% (27)
	Agreement	Disagreement	Agreement	Disagreement
R1-S	81% (103) $\kappa = 0.51$	19% (24)	87% (110) $\kappa = 0.58$	13% (17)
R2-S	73% (93) $\kappa = 0.15$	27% (34)	82% (104) $\kappa = 0.18$	18% (23)
R1-R2	81% (103) $\kappa = 0.25$		86% (109) $\kappa = 0.35$	

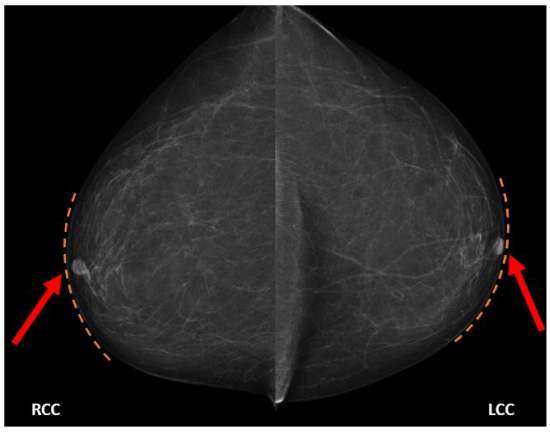

Figure 9: Nipple is not in profile

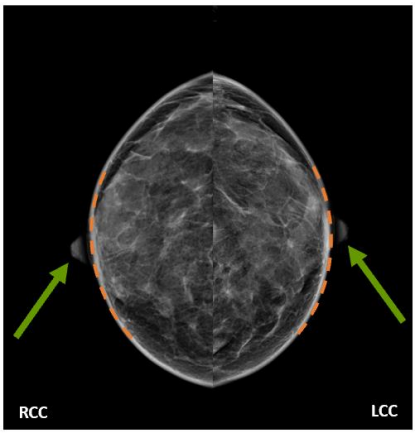

Figure 10: Nipple is in profile

6.2.5 Sufficient breast tissue at medial and lateral side

The percentage of images with "Tissue cut off" or in other words "missed tissue at medial and/or lateral side of the breast" was 8%, 36% and 3% for the radiographer, radiologist and software in CC view and 2%, 20% and 0% in MLO view respectively (table 8). The agreement for tissue cut off in CC view is 92% and 62% for radiographer and software and for radiologist and software respectively. The kappa agreement was only fair for radiographer and software ($\kappa = 0.25$) and poor for radiologist and software ($\kappa = -0.019$). For MLO view the agreement between radiographer and software was 98% and between radiologist and software it was 80%. The kappa agreement could not be calculated for MLO view since the software scored no image with tissue cut off. The kappa values between readers is 0.14 and - 0.03 for CC and MLO view resp. Figure 11 and 12 are examples of tissue cut off in CC and MLO view resp. In the first example the radiologist did not agree with the radiographer and software as she indicates that there is no cut off, in contrast to the second example where the radiologist is the only one who assessed the cut off correctly.

Table 8: Percentage (%) and proportion (n = out of 127 examinations) of tissue cut off for CC and MLO view by the radiographer (R1), radiologist (R2) and Software (S), and the agreement, disagreement between radiographer and Volpara (R1- S) and radiologist and Volpara (R2-S)

	Tissue Cut Off – CC view		Tissue Cut Off – MLO view	
	Yes	No	Yes	No
	% (n=)	% (n=)	% (n=)	% (n=)
R1	8% (10)	92% (117)	2% (2)	98% (125)
R2	36% (46)	64% (81)	20% (25)	80% (102)
S	3% (4)	97% (123)	0% (0)	100% (127)
	Agreement	Disagreement	Agreement	Disagreement
R1-S	92% (117) $\kappa = 0.25$	8% (10)	98% (125) κ = NM	2% (2)
R2-S	62% (79) $\kappa = -0.019$	38% (48)	80% (102) κ = NM	20% (25)
R1-R2	67% (85) $\kappa = 0.14$		79% (100) $\kappa = -0.030$	

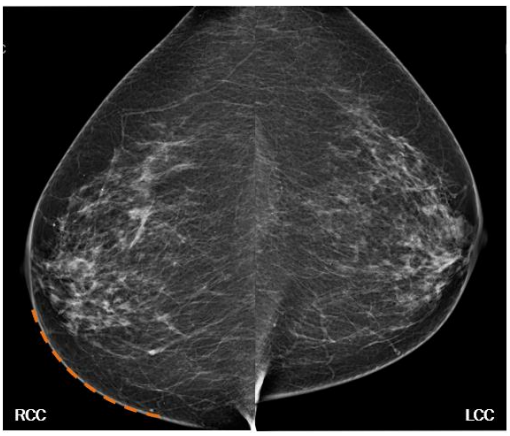

Figure 11: Tissue cut off on the medial side of the breast in right CC view.

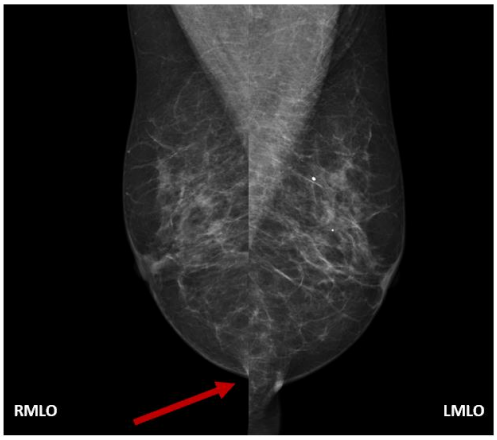

Figure 12: Tissue cut off in right MLO view.

6.2.6 Inframammary fold on MLO view

The percentage of images in which the inframammary fold was visible, was 43%, 37% and 35% for the radiographer, radiologist and software respectively. For this metric the software was more severe than the readers. The inframammary angle was obscured in 32%, 33% and 43% of all images resp. and in 26%, 31% and 22% resp. the inframammary fold was absent (table 9). The kappa value for radiographer and software was 0.57, and 0.53 for the radiologist and software. Figure 13 gives an example of this metric. The software stated that in the left image, the IMF is absent. The middle image shows an obscured IMF. In the right image, the IMF is visible. On the other hand the radiologist assessed the second example as 'visible' instead of 'obscured' like the radiographer and software did.

Table 9: Percentage (%) and proportion (n = out of 127 examinations) of inframammary angle depiction for MLO view by the radiographer (R1), radiologist (R2) and Software (S), and the agreement, disagreement between radiographer and Volpara (R1- S) and radiologist and Volpara (R2-S)

	Inframammary angle depiction		
	Visible	Obscured	Absent
	% (n=)	% (n=)	% (n=)
R1	43% (54)	32% (40)	26% (33)
R2	37% (47)	33% (41)	31% (39)
S	35% (45)	43% (54)	22% (28)
	Agreement	Disagreement	
R1-S	72% (91)	28% (36)	
	$\kappa = 0.57$		
R2-S	69% (87)	32% (40)	
	$\kappa = 0.53$		
R1-R2	69% (88)		
	$\kappa = 0.57$		

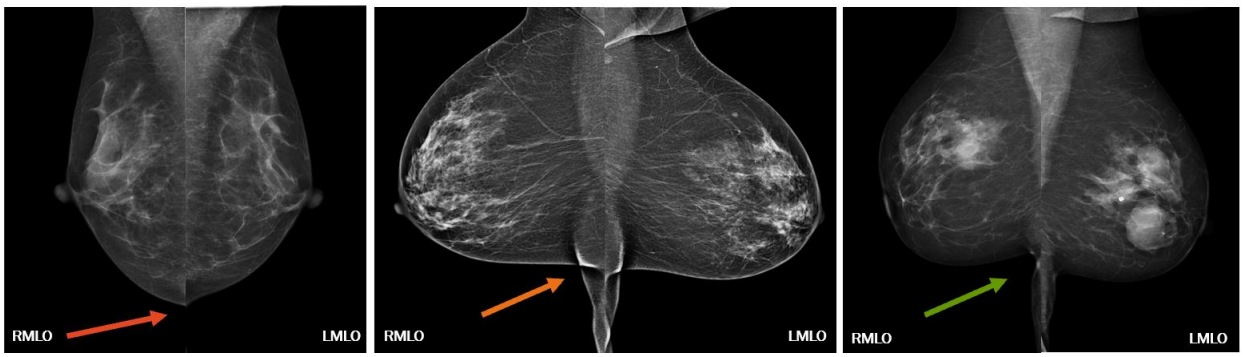

Figure 13: Example of the different outcomes for the inframammary angle

6.2.7 Skin folds in the MLO image

In table 10, the percentage of MLO images where skin folds were present within the pectoral muscle are tabulated and this was the case in 34%, 47% and 39% of the cases for the radiographer, radiologist and software resp. The kappa between the radiographer and software was 0.15 and between the radiologist and software 0.36. In 40% of the cases the radiographer noted a skin fold in the breast tissue and this was in 12% of the cases for the radiologist. It is possible that a fold is present in both pectoral and breast, but readers could not indicate both and thus had to outweigh which fold was more pronounced. The folds in breast tissue scored by the readers could not be compared with the software as it is not yet possible for the software to detect this. Figure 14 and 15 are examples skin folds in breast images that were detected by the TruPGMI software as well as the readers.

Table 10: Percentage (%) and proportion (n = out of 127 examinations) of pectoral folds for MLO view by the radiographer (R1), radiologist (R2) and Software (S), and the agreement, disagreement between radiographer and Volpara (R1-S) and radiologist and Volpara (R2-S)

	MLO folds	
	Yes	No
	% (n=)	% (n=)
R1	34% (43)	61% (84)
R2	47% (59)	54% (68)
S	39% (49)	61% (78)
	Agreement	Disagreement
R1-S	61% (77) $\kappa = 0.15$	39% (50)
R2-S	69% (87) $\kappa = 0.36$	32% (40)
R1-R2	73% (93) $\kappa = 0.45$	27% (34)

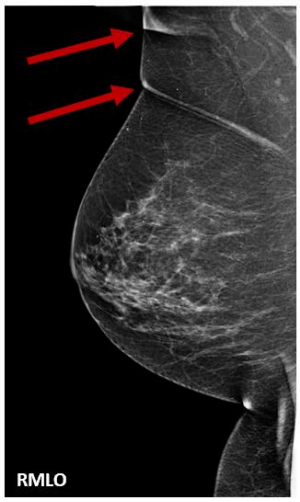

Figure 14: Folds in breast tissue

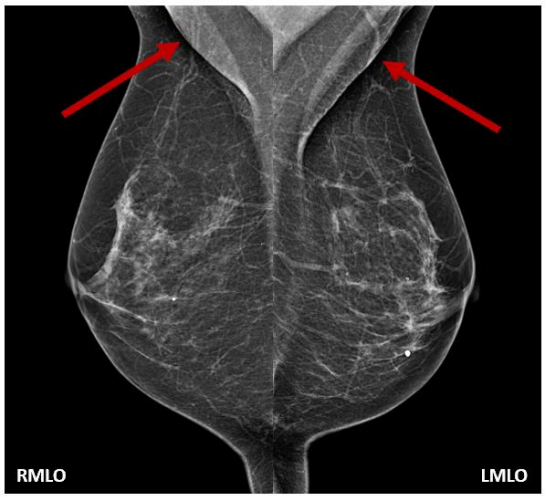

Figure 15: Folds in pectoral muscle

6.2.8 Adequate depiction of the pectoral muscle in MLO view

The percentage of images that have sufficiently deep depiction of the pectoral muscle was 88%, 73% and 96% for the radiographer, radiologist and software resp. In this work, a similar number of images had a pectoral muscle which was acquired sufficiently deep according to the readers. Kappa values of 0.30 and 0.19 were found for the readers versus the software. The percentage of images that were rated sufficiently wide was 89%, 54% and 91% for the radiographer, radiologist and software resp. For both metrics depth and width, the radiologist was more severe than the radiographer and the software. The agreement between the radiographer and radiologist was 83% and 63% for depth and width resp. with kappa values of 0.49 and 0.22 (table 11). In figure 16 the total posterior breast length is indicated together with the posterior pectoral muscle length (orange dotted lines). The length of the pectoral muscle should be at least one-third of the full height of the breast. This is the case for the right breast but the left breast is too short. Figure 17 represents the adequate width for the pectoral muscle where it should lay between indicated threshold. For both examples, the readers and software agreed on this.

Table 11: Percentage (%) and proportion (n = out of 127 examinations) of pectoral adequacy for MLO view by the radiographer (R1), radiologist (R2) and Software (S), and the agreement, disagreement between radiographer and Volpara (R1-S) and radiologist and Volpara (R2-S)

	Sufficiently deep		Sufficiently wide	
	Yes	No	Yes	No
	% (n=)	% (n=)	% (n=)	% (n=)
R1	88% (112)	12% (15)	89% (113)	11% (14)
R2	73% (93)	27% (34)	54% (68)	47% (59)
S	96% (122)	4% (5)	91% (116)	9% (11)
	Agreement	Disagreement	Agreement	Disagreement
R1-S	84% (107) $\kappa = 0.30$	16% (20)	84% (107) $\kappa = 0.26$	16% (20)
R2-S	72% (91) $\kappa = 0.19$	28% (36)	57% (72) $\kappa = 0.12$	43% (55)
R1-R2	83% (106) $\kappa = 0.49$		63% (80) $\kappa = 0.22$	

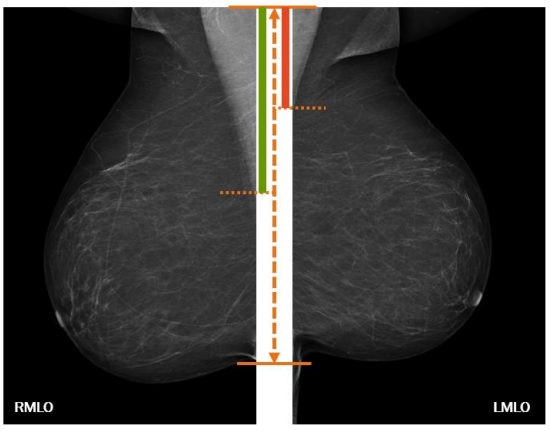

Figure 16: Adequate depth of the pectoral muscle in MLO view

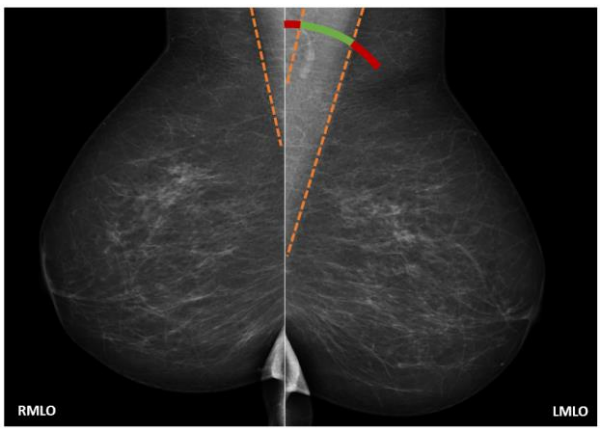

Figure 17: Adequate width of the pectoral muscle in MLO view

6.2.9 Overall quality score

When looking at the overall quality of the CC view, one can conclude that a limited number of images fall into the categories Inadequate and Perfect. Only the radiographer scored more images as Perfect than the radiologist and software, 14% (R1) vs 4% (R2) and 2% (S). The vast majority of images are present in the Good and Moderate group. The radiographer and the radiologist concluded that 62% and 65% of the images could be categorised as Good. This while the software only stated this for 11% of the images. On the other hand, the software included the vast majority in the Moderate (84%) category. A very low agreement is observed both for the radiographer (24%) and the radiologist (32%) for CC view. The agreement for MLO view between the radiographer and software was 54% and 50% between radiologist and software. Again, a lower percentage of the categories Perfect and Inadequate was seen (table 12).

Table 12: Percentage (%) and proportion (n = out of 127 examinations) of the overall quality for CC and MLO view by the radiographer (R1), radiologist (R2) and Software (S), and the agreement, disagreement between radiographer and Volpara (R1-S) and radiologist and Volpara (R2-S)

	Overall quality – CC view				Overall quality – MLO view			
	P	G	M	I	P	G	M	I
	% (n=)	% (n=)	% (n=)	% (n=)	% (n=)	% (n=)	% (n=)	% (n=)
R1	14% (18)	62% (79)	21% (26)	3% (4)	6% (8)	63% (80)	24% (31)	6% (8)
R2	4% (5)	65% (82)	30% (38)	2% (2)	4% (5)	54% (69)	38% (48)	0
S	2% (3)	11% (14)	84% (106)	3% (4)	8% (10)	57% (72)	29% (37)	6% (8)
	Agreement	Disagreement			Agreement	Disagreement		
R1-S	24% (30)	76% (97)			54% (69)	46% (58)		
	κ = - 0.10				κ = 0.19			
R2-S	32% (41)	68% (86)			50% (63)	50% (64)		
	κ = 0.0010				κ = 0.13			
R1-R2	46% (59)				44% (56)			
	κ = 0.020				κ = -0.022			

When one combines the categories Perfect, Good and Moderate together such that a binary classification is obtained, the agreement increased to 97% for both CC and MLO view between radiographer and software and to 95% and 93% for CC and MLO respectively between radiologist and software (table 13). The kappa values for the radiographer and software increased, in going from - 0.10 to 0.484 for CC view and from 0.19 to 0.733 for MLO view. On the other hand ,the kappa agreement between the radiologist and software decreased to -0.201 and -0.051 for CC and MLO view resp.

Table 13: Agreement (%) and kappa between radiographer (R1) and radiologist (R2) vs software for the PGMI score compared to the binary PGM vs I score

		Binary overall quality – CC		Binary overall quality – MLO	
		PGMI	PGM vs I	PGMI	PGM vs I
R1-S	κ_w	- 0.10	0.484	0.19	0.733
	Agreement	24%	97%	54%	97%
R2-S	κ_w	0.001	-0.021	0.13	-0.051
	Agreement	32%	95%	50%	93%

6.2.10 Image retake

The percentage of images that needed retake was 8%, 7% and 3% for CC view for the radiographer, radiologist and the software respectively. For the MLO view, both readers advised a retake in 12% of the images. The software does not aim to decide for retakes. The software leaves this decision at the discretion of the radiologist. However we have used the binary score of overall positioning quality where we interpreted the inadequate score as an advice for retake. However we would like to note that this is not the way the software is intended to be used. The software gave an inadequate score in 6% of all images. The agreement between the radiographer and the software was 94% for CC view and 93% for MLO view, and 90% and 88% resp. for the radiologist and the software. However there were only 4 cases that were assessed for retake by both the radiologist and the software, and only 10 cases by both the radiographer and the software. This resulted in an inter-rater reliability for the radiologist and radiographer versus the software of -0.05 and 0.40 for CC view respectively and 0.29 and 0.57 for MLO view resp. (table 14).

Table 14: Percentage (%) and proportion (n = out of 127 examinations) of advised retakes for CC and MLO view by the radiographer (R1), radiologist (R2) and Software (S), and the agreement, disagreement between radiographer and Volpara (R1- S) and radiologist and Volpara (R2-S)

	Retake – CC view		Retake – MLO view	
	Yes	No	Yes	No
	% (n=)	% (n=)	% (n=)	% (n=)
R1	8% (10)	92% (117)	12% (15)	88% (122)
R2	7% (9)	93% (118)	12% (15)	88% (112)
S	3% (4)	97% (123)	6% (8)	94% (119)
	Agreement	Disagreement	Agreement	Disagreement
R1-S	94% (119) $\kappa = 0.40$	6% (8)	93% (118) $\kappa = 0.57$	7% (9)
R2-S	90% (114) $\kappa = -0.046$	10% (13)	88% (112) $\kappa = 0.29$	12% (15)
R1-R2	88% (112) $\kappa = 0.15$		87% (111) $\kappa = 0.40$	

Figure 18 and 19 represent the number of images that required a retake according to the radiologist and radiographer. Not all of these images were scored as inadequate for their overall quality, a large portion received a quality score of moderate. For the advised retakes in CC view by the radiographer, 4 of these images were inadequate and 6 moderate. The radiologist concluded 2 as inadequate and 7 as moderate. In MLO view there were more images that were considered for retake than in CC view. The radiographer scored 8 retake images as inadequate and 7 as moderate in MLO view. The radiologist concluded 5 retake images as inadequate and 10 as moderate. These data show us that an overall quality score as provided by the TruPGMI software does not say it all. The decision for retake is made after careful consideration of many other aspects where the patient is central.

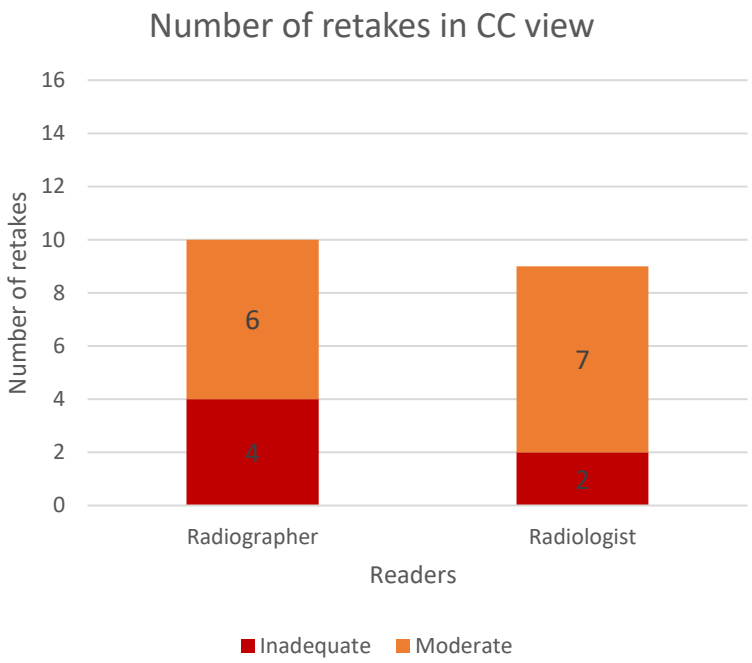

Figure 18: Histogram of the number of CC-images considered for retake with their overall quality score

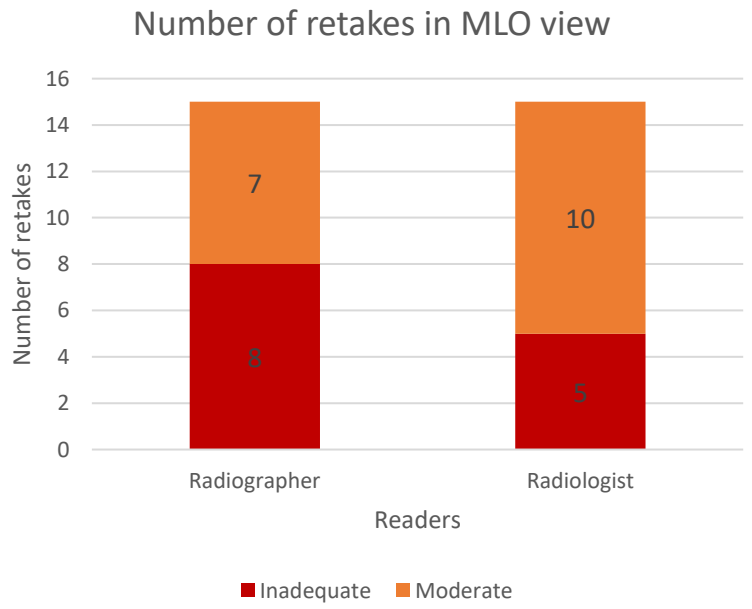

Figure 19: Histogram of the number of MLO-images considered for retake with their overall quality score

6.3 Reproducibility testing using different vendors

6.3.1 Nipple in profile

Table 15 shows the outcome of the software for nipple in profile for CC and MLO view for the different vendors of mammographic systems. A p-value of <.05 is considered significant and when a paired wise comparison between the three vendors is done, a p-value of <.017 is considered significant. A chi-square test of independence for CC view showed that there was no significant difference between the distributions of 'yes' and 'no' between the different vendor image , χ^2 (2, N = 526) = 4.1, p = .13. For MLO view the chi-square showed a significant difference, χ^2 (2, N = 536) = 9.1, p = .01. Further analysis with a post-hoc Fisher exact test showed no significant difference between GE and Hologic, but a significant difference did exist between Siemens and Hologic (p = .007) and between Siemens and GE (p = .006) for MLO view.

Table 15: Percentage (%) and proportion (n = out of x examinations) of nipple in profile for both CC and MLO view for GE, Hologic and Siemens

	Nipple in profile – CC view			Nipple in profile – MLO view		
	GE	Hologic	Siemens	GE	Hologic	Siemens
	% (n=186)	% (n=213)	% (n=127)	% (n=186)	% (n=213)	% (n=127)
Yes	71% (132)	62% (132)	69% (88)	64% (119)	65% (138)	79% (100)
No	29% (54)	38% (81)	31% (39)	36% (67)	35% (75)	21% (27)

6.3.2 Tissue cut off

A chi-square test could not be performed for the criterion 'tissue cut off' because this test cannot handle zeros in the contingency table (table 16). Instead, a Fisher test is performed between the unpaired scoring of presence or absence of tissue cut off. A significant difference is only seen for CC view between GE and Hologic (p = .001). All outcomes for tissue cut off in MLO view are the same, as the software saw no tissue cut off in any of the images. Therefore the association is considered to be not statistically significant (p = 1).

Table 16: Percentage (%) and proportion (n = out of x examinations) of tissue cut off for both CC and MLO view for GE, Hologic and Siemens

	Tissue Cut Off – CC view			Tissue Cut Off – MLO view		
	GE	Hologic	Siemens	GE	Hologic	Siemens
	% (n=186)	% (n=213)	% (n=127)	% (n=186)	% (n=213)	% (n=127)
Yes	0% (0)	5% (11)	3% (4)	0% (0)	0% (0)	0% (0)
No	100% (186)	95% (202)	97% (123)	100% (186)	100% (213)	100% (127)

6.3.3 Inframammary angle

For the criterion 'inframammary angle depiction' the chi-square test showed significant differences between the scores visible, obscured or absent for the different vendor images, χ^2 (4, N = 536) = 17.6, p = .002 (table 17). A second chi-square was done to make a paired wise comparison instead of the Fisher exact test. This was done so because the used calculator for a 3x3 contingency table test could not handle integer numbers above 20. A significant difference is only seen between GE and Siemens with a p-value of less than 0.001.

Table 17: Percentage (%) and proportion (n = out of x examinations) of inframammary angle depiction for MLO view for GE, Hologic and Siemens

	Inframammary angle depiction		
	GE	Hologic	Siemens
	% (n=186)	% (n=213)	% (n=127)
Visible	33% (62)	34% (72)	35% (45)
Obscured	60% (111)	54% (114)	43% (54)
Absent	7% (13)	13% (27)	22% (28)

6.3.4 Pectoral adequacy

There was no statistical association for the criterion 'pectoral sufficiently deep', χ^2 (2, N = 526) = 0.67, p = .71. For the criterion 'pectoral sufficiently wide' the chi-square test did show a statistically significant difference in scoring the pectoral adequacy, χ^2 (2, N = 526) = 9.1, p = .01. The Fisher's exact test showed that this difference was present between GE and Hologic (p = .008) and between GE and Siemens (p = .004). The p-value is .7 for Siemens and Hologic and therefore there is no statistically significant difference between the two (table 18).

Table 18: Percentage (%) and proportion (n = out of x examinations) of pectoral adequacy for MLO view for GE, Hologic and Siemens

	Sufficiently deep			Sufficiently wide		
	GE	Hologic	Siemens	GE	Hologic	Siemens
	% (n=186)	% (n=213)	% (n=127)	% (n=186)	% (n=213)	% (n=127)
Yes	96% (178)	97% (207)	96% (122)	98% (183)	92% (197)	91% (116)
No	4% (8)	3% (6)	4% (5)	2% (3)	8% (16)	9% (11)

6.3.5 Folds in MLO view

Table 19 shows the results for the criterion 'MLO folds' for the different vendors. There was a significant difference in scoring if folds were present or not between the different vendors, χ^2 (2, N=526) = 23.1, p < .0001. The two-tailed p-value of the Fisher exact test between Siemens and Hologic p = .41, which implies no statistical significance between the two. Both between GE and Siemens and between GE and Hologic have a p-value less than .0001.

Table 19: Percentage (%) and proportion (n = out of x examinations) of folds in MLO view for GE, Hologic and Siemens

	MLO folds		
	Ge	Hologic	Siemens
	% (n=186)	% (n=213)	% (n=127)
Yes	16% (30)	34% (72)	39% (49)
No	84% (156)	66% (141)	61% (78)

6.3.6 Overall quality

For the overall quality in CC view a chi-square test was performed. However, a zero is present in the contingency table, therefore the statistical result is less reliable. Nevertheless, a significant difference is found in CC view ($p < .001$). The difference was found between GE and Hologic ($p < .001$) and between GE and Siemens ($p < .001$). We did a small check if we can rely on this statistics outcome: when removing the data row of the "inadequate" category so that the contingency table does not contain any zero, we still find a significant difference between these pairs. A chi-square test of independence showed that there was no significant difference between the vendors and the overall quality scoring in MLO view, χ^2 (6, N = 526) = 2.99, p = .81 (table 20).

Table 20: Percentage (%) and proportion (n = out of x examinations) of overall quality for both CC and MLO view for GE, Hologic and Siemens

	Overall quality – CC view			Overall quality – MLO view		
	GE	Hologic	Siemens	GE	Hologic	Siemens
	% (n=186)	% (n=213)	% (n=127)	% (n=186)	% (n=213)	% (n=127)
P	11% (21)	5% (11)	2% (3)	6% (12)	8% (18)	8% (10)
G	33% (62)	17% (36)	11% (14)	59% (110)	62% (132)	57% (72)
M	55% (103)	73% (155)	84% (106)	30% (55)	26% (56)	29% (37)
I	0% (0)	5% (11)	3% (4)	5% (9)	3% (7)	6% (8)

6.4 The influence of other parameters

An analysis has been performed on the influence of different independent parameters on positioning quality through ordinal regression. It showed that the volumetric breast density class (VDG) and the breast volume did not have an influence on the overall quality. The influence of the compression force was statistically significant ($p <.001$). The estimate is the ordered log odds of the impact of that specific parameter on the overall quality. To know the proportional odds ratio, the exponent of the estimate needs to be calculated (table 21). For a one unit increase in compression force, we expect a 1.015 increase in the ordered odds of being in a higher level of PGMI, given all of the other variables are held constant. The same was observed for compression pressure ($p < .001$) where a 0.948 increase in PGMI level is seen for a one unit increase in compression pressure. The most significant influence was observed for the dose ($p < .001$). For a one unit increase in dose, we expect a 1.675 increase in the ordered odds of being in a higher level of PGMI, given all the other variables are held constant. The average dose in the dataset used for this regression was 1,256 mGy.

Table 21: Estimate and probability outcomes of the SPSS analysis for the ordinal regression on the different parameters with the calculated exponent of the estimate for the significant parameters

	Estimate	**Exp(Estimate)**	**Probability**
Breast volume	7.5E-5		.175
Compression Pressure	-0.53	0.948	< .001
Compression Force	0.15	1.015	< .001
Dose	0.516	1.675	< .001
VDG			
a	-0.094		.385
b	0.023		.806
c	0.125		.202
d	0		/

7 Discussion

This study reports the objective versus subjective evaluation of the positioning quality of breast cancer screening mammograms. A reading study was conducted comparing the software output with the scores of an experienced radiologist and radiographer. When looking at the agreements in percentage and kappa agreement, different conclusions can be drawn.

Criteria 1 and 2 on the depiction of skin folds in the breast tissue and the pectoral muscle posterior in the CC projection view as well as the symmetry in CC and MLO view were not included in the software and therefore no comparison could be made. For the depiction of the pectoral muscle in CC view, a slight hint of the pectoralis muscle in the image resulted in a score 'Good' for the radiologist whereas only a score 'Moderate' was given by the radiographer. It needs to be noted that the depiction of the pectoral muscle shown at the posterior edge of the breast in CC view is not consistently achieved as it depends on anatomical characteristics of the woman. Also van Landsveld-Verhoeven et al. and Sweeney et al. reported low numbers of CC images where the pectoral muscle was visible [3, 16]. Measurements for determining symmetry between left and right breast do not yet exist in the Volpara TruPGMI software. However, symmetry could possibly be derived from metrics such as the posterior breast length or posterior nipple length.

The percentage of response of the metric 'nipple in profile' are slightly higher to those found by Waade et al. where 76% to 85% of the images fulfilled the nipple in profile criterion [20]. Based on our current assessment one can conclude that the agreement values for the nipple in profile criterion are sufficiently high but still slightly lower than those found by Waade et al. with percentage agreements ranging from 90% to 96% [20]. The kappa values for nipple in profile for CC and MLO view of the radiographer and radiologist vs software are considerably lower than the ones ($\kappa > 0.69$) published in Waade et al. [20]. A better correspondence between the radiographer and the software was obtained compared to the scores of the radiologist, resulting in only a fair agreement between the two readers ($\kappa = 0.25$ and 0.35 for CC and MLO resp.). Visual inspection of the images showed that the software had more difficulties to assess a subtle presence of the nipple, and to discriminate the nipple with skin. The readers were more severe compared to the software.

The criterion for missed tissue in CC view were in fact two separate questions in the reader study (question 3 and 4 in table 4). Therefore the software score "yes" was in agreement when the readers scored "yes" to question 3 (lateral side) OR "yes" to question 4 (medial side). The software as well as the radiographer often claimed there is no missing tissue in contradiction to the radiologist, leading to a poor agreement between the two readers ($\kappa = 0.14$ and -0.03 for CC and MLO resp.). A second look to these images by a third radiologist showed that the missed tissue was not only scored for the medial and lateral side but also the missed tissue in the anterior-posterior direction was considered. After all, in the Flemish screening the evaluation criterion of missed tissue includes all sides of the complete breast. Overall, it was observed that missed tissue in the images occurred more often in CC than in MLO view.

For the metric of the inframammary fold visibility in MLO view, the percentage of images with visible inframammary fold, the software was more severe than the readers. This is opposite to what is described in Waade et al. where the software indicates good visibility of the inframammary fold in 54% of the cases compared to a lower percentage of 39% and 47% by the radiographers [20]. The number of cases where the inframammary fold was not included in the images was lower in the dataset of Waade et al. [20]. In principle, one expects that the inframammary fold or at least a hint of the fold is

visible in order to obtain sufficient breast tissue on the image. Our results suggest that it is very common that the inframammary fold is obscured by skin folds, but this is not always recognized as a problem for the diagnostic quality.

Both the results of the software and the readers suggest that the number of skin folds in the cases are higher that one would optimally want. The agreement levels between the readers and the software are in line with published data of Waade et al. with kappa values of 0.18 and 0.28 for the radiographers versus software [20]. In the reader study, the readers had the possibility to answer where the fold was seen, in the pectoral muscle or in the breast tissue. When processing the results it became clear that when the radiologist answered 'breast', it is not possible to compare this outcome with the software as the software cannot recognize skin folds in the breast tissue but only in the pectoral muscle.

In the software "Pectoral adequacy" is evaluated based on 2 metrics, if the muscle is sufficiently deep (towards the nipple) or sufficiently wide presented on the images. The width of the pectoral muscle can be rated as narrow, adequate or too wide, where narrow and wide should both be interpreted as suboptimal. The depth of the pectoral muscle was rated as adequate or too short. The software scored remarkably higher compared to the readers on the question about the adequate depth depiction of the pectoral muscle. The opposite was found in Waade et al. with only 62% of the cases for which the pectoral muscle reached 1 cm or more below pectoral nipple line according to the software, compared to 71% and 76% according to the radiographers. In this work, a similar number of images had a pectoral muscle which was acquired sufficiently deep according to the readers. For both metrics depth and width, the radiologist was more severe than the radiographer and the software.

An overall image score in the form of a PGMI score is derived out of the previous criteria described in tables 5-14. The goal of monitoring positioning quality is to minimise the Moderate and Inadequate categorized images. A very low agreement was observed both for the radiographer and the radiologist for CC view. The agreement for MLO view between the radiographer and software was almost double compared to the agreement for CC images. A possible explanation for this low agreement in the overall quality is because of the four nominal levels of Perfect, Good, Moderate and Inadequate and the fact that the readers did not have the same threshold for each quality level. The radiologist scored more often moderate while the technologist gave more often the score good.

In order to reduce the effect of the four thresholds for quality scoring, the ratings were changed to a binary score (Perfect, Good, Moderate versus Inadequate) (table 15). One could see that the agreement becomes almost 100% for both views and both readers vs software. When looking closer to the 3% disagreement between radiographer and software in CC and MLO view and 5% and 7% disagreement between radiologist and software for CC and for MLO view resp., these disagreements included only the inadequate cases. This is even better represented by the kappa values for the binary PGM vs I score. It is noteworthy that the kappa value between radiologist and software for both CC and MLO view decrease significantly. Because the kappa test takes into account not only the agreements but also the discordant cases, this statistical test is better for the interpretation of these results. Further analysing these kappa values with the cross tables generated by SPSS, one could see that there was no agreement on the inadequate cases between radiologist and software both for CC and MLO view for the binary PGM vs I score. Conversely, we see that the kappa value between radiographer and software for the binary PGM vs I score increases, from no agreement (κ_w = -0.1) to moderate agreement (κ_w = 0.484) in CC view and from slight agreement (κ_w =0.19) to substantial agreement (κ_w =0.733) in MLO view. The concordance of the assessment of inadequate cases which might lead to retake is the most important outcome of this software.

Next to an overall positioning quality score following the PGMI metric, the readers were also asked to identify images that needed retake. As expected, the total percentage of advised retakes is low**.** Hence, the European guidelines for breast cancer screening have set a lower limit of 85% for the number of mammograms that should achieve a good quality and the recall rate should be below 5% at initial screening [1]. In this study a difference in number of retakes between CC and MLO view was found; in MLO view the number of advised retakes was double than in CC view. This is probably due to the larger number of criteria to meet and the fact that the breast is more difficult to position in MLO projection. As mentioned above, in the reader study, a distinction was made between the question about retake and inadequate images in the PGMI score for determining overall quality. In the software, this is not done and there is no decision if a retake is needed or not. The software leaves this decision at the discretion of the radiologist. One has therefore interpreted the inadequate image score as a retake. When readers score an image as inadequate they also designate it as retake but vice versa this is not always true. It is seen that a retake image is also scored as moderate (figure 13 and 14). Beside the positioning quality, a retake can be considered for technical reasons and should be kept under 5%, ideally 3% [1]. Sometimes the patient is unable to be better positioned and, despite an inadequate image, a repeat may not be useful. Thus, it is clear that the decision to perform a retake has several aspects that must be taken into account and it can only be made based on interaction with the radiologist.

To assess the reproducibility of the software, different vendors of mammographic systems were included. Each criteria is again evaluated with the software. For each criterion, a statistically significant difference was found between some vendors, differences appeared most often between GE and Hologic and GE and Siemens. Between Siemens and Hologic, only the nipple in profile criterion scored differently. The problem with this reproducibility test is that the datasets of the different vendor systems were also coming from different radiology practices and therefore different technologists doing the positioning. It remains to be tested if the differences that were found were merely due to the variation in systems or if actual positioning quality differences were inherent between the datasets. A reader study for a subset of patient mammograms for each system may help in clarifying this and is part of future work. In addition the expansion of the datasets including more images may also help in this regard.

Through ordinal regression, the possible influence of patient specific or acquisition parameters on the positioning quality has been assessed. Compression force as well as compression pressure have an impact on the positioning quality, which was expected. Compression pressure is the compression force applied by the paddle, divided by the area of the breast. So for a small or large breast, where the same compression force is applied, different compression pressures are experienced. Holland et al. showed that women with the lowest received compression pressure are more often recalled [16]. Also in this study, one sees that with higher compression pressure and force, there is a better positioning quality score. The dose had the largest influence on positioning quality, which is a surprising result. Good positioning may result in an optimal selection of the automatic exposure control cell and result in an optimally set dose, however this is therefore not the highest dose [1, 21].

This study has several limitations. First, to our knowledge, only one automatic evaluation of mammography positioning software is available and was tested in this study. Second, the number of readers was limited to two, only one radiographer and one radiologist participated. A second radiologist has provided additional feedback for discordant cases. Third, the number of mammograms evaluated in the reading study was also limited. However the number of quality criteria was rather extended and increased the reading time considerably. A last limitation is that the software was not yet calibrated for our hospital's interpretation of positioning requirements.

8 Conclusion

This study has demonstrated that for the monitoring of positioning of mammographic screenings exams based on quality criteria, differences can be observed between software (objective assessment) and readers (subjective assessment). Due to the setup of the reader study or the choice of the quality criteria, it was not possible to compare all of the results with the Volpara software, such as folds and the pectoral depiction in the CC view. For the final overall positioning quality assessment the readers and software showed good agreement.

The radiographer had an overall good agreement with the software compared to the radiologist who was less in line with the software. Seeing this better agreement between the radiographer who is our local teacher in screening mammograms and the software, one can conclude that the software would be useful as a continues training tool, providing immediate feedback to the radiographers.

Of all criteria, it is most important to agree on the cases that are scored inadequate as these might result in retake based on the radiologist's final interpretation. The retakes and inadequate cases were assessed separately in the readers study but in the software only inadequate cases were assessed and so interpreted as retake. This difference must be handled with care. This agreement on inadequate images was found to be moderate to high between radiographer and software but there was no agreement between radiologist and software.

In addition to the actions of the technologist to position the breast correctly, the compression force and pressure along with the dose also affect the positioning quality. Therefore, this should also be taken into account during positioning.

Subjective positioning quality monitoring is prone to high reader variability; this can be overcome via the use of automatic measurements with software.

Prior to the use of any software for medical use, a careful validation is needed. The purpose of this thesis was to perform this validation of the automatic quality monitoring software using objective tests of accuracy, reproducibility, and robustness. This validation and evaluation was successfully completed.

References

[1] N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. van Karsa, *European guidelines for quality assurance in breast cancer screening and diagnosis*. Luxembourg: Office for Official

Publications of the European Communities, 2006.

[2] CvKO. "Bevolkingsonderzoek borstkanker." https://borstkanker.bevolkingsonderzoek.be/nl/bk/wat-het-bevolkingsonderzoek-borstkanker (accessed 14/04, 2022).

[3] C. van Landsveld-Verhoeven, G. J. den Heeten, J. Timmers, and M. J. Broeders, "Mammographic positioning quality of newly trained versus experienced radiographers in the Dutch breast cancer screening programme," *Eur Radiol,* vol. 25, no. 11, pp. 3322-7, Nov 2015, doi: 10.1007/s00330-015-3738-8.

[4] Vlaanderen. "Zorg en gezondheid Vlaanderen." https://www.zorg-en-gezondheid.be/bevolkingsonderzoek (accessed 15/04, 2022).

[5] D. Paulus, F. Mambourg, and L. Bonneux, "Borstkankerscreening," Federaal Kenniscentrum voor de Gezondheidszorg, 2005.

[6] IAEA, "Quality Assurance Programme for Digital Mammography," vol. 17. Austria, 2011, ch. 1-2, pp. 1-16.

[7] "Belgium: Females, age-specific and age-standardised incidence rates of cancer, by primary site in 2019 (n/100,000 person years)," ed, 2019.

[8] "Cancer Fact Sheet Breast Cancer," Belgian cancer register, 2019. [Online]. Available: https://kankerregister.org/media/docs/CancerFactSheets/2019/Cancer_Fact_Sheet_FemaleBreastCancer_2019.pdf

[9] "BELGISCHE KANKERBAROMETER Editie 2021 Hoofdstuk 3 - Screening," Stichting tegen Kanker, Brussel, 2021.

[10] *Besluit van de Vlaamse REgering betreffende bevolkingsonderzoek in het kader van ziektepreventie*, 2008.

[11] FANC. "Jaarlijkse gemiddelde blootstelling aan ioniserende straling in België." https://fanc.fgov.be/nl/informatiedossiers/wat-radioactiviteit-ioniserende-straling/gemiddelde-blootstelling (accessed 4 april, 2022).

[12] "Vlaams draaiboek bevolkingsonderzoeken naar kanker," 2021.

[13] M. Goossens *et al.*, "Flemish breast cancer screening programme: 15 years of key performance indicators (2002-2016)," *BMC Cancer,* vol. 19, no. 1, p. 1012, Oct 28 2019, doi: 10.1186/s12885-019-6230-z.

[14] "Jaarrapport 2021 BVO naar kanker," Bevolkingsgroep.be, 2021.

[15] L. R. v. Bevolkingsonderzoek, *Syllabus: Fysische aspecten bij digitale mammografie t.b.v. de laboranten van het Nederlandse Bevolkingsonderzoek op Borstkanker*. 2014.

[16] K. Holland, I. Sechopoulos, R. M. Mann, G. J. den Heeten, C. H. van Gils, and N. Karssemeijer, "Influence of breast compression pressure on the performance of population-based mammography screening," *Breast Cancer Res,* vol. 19, no. 1, p. 126, Nov 28 2017, doi: 10.1186/s13058-017-0917-3.

[17] A. Svalkvist, S. Svensson, T. Hagberg, and M. Bath, "Viewdex 3.0-Recent Development of a Software Application Facilitating Assessment of Image Quality and Observer Performance," *Radiat Prot Dosimetry,* vol. 195, no. 3-4, pp. 372-377, Oct 12 2021, doi: 10.1093/rpd/ncab014.

[18] A. Svalkvist, S. Svensson, M. Håkansson, M. Båth, and L. G. Månsson, "ViewDEX:a status report," *Radiat. Prot. Dosimetry,* vol. 169, no. 1-4, pp. 38-45, 2016.

[19] K. A. Hallgren, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," *Tutor Quant Methods Psychol,* vol. 8, no. 1, pp. 23–34, 2012.

[20] G. G. Waade *et al.*, "Assessment of breast positioning criteria in mammographic screening: Agreement between artificial intelligence software and radiographers," *J Med Screen,* vol. 28, no. 4, pp. 448-455, Dec 2021, doi: 10.1177/0969141321998718.
[21] C. J. Martin and D. G. Sutton, "Radiation Dose and Image Quality," *Cancer Imaging: Lung and Breast Carcinomas,* pp. 45-62, 2008.
[22] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem Med (Zagreb),* vol. 22, no. 3, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/?report=classic.
[23] M. F. Zibran, "CHI-Squared Test of Independence."
[24] M. L. McHugh, "The chi-square test of independence," *Biochem Med (Zagreb),* vol. 23, no. 2, pp. 143-9, 2013, doi: 10.11613/bm.2013.018.
[25] H. Y. Kim, "Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test," *Restor Dent Endod,* vol. 42, no. 2, pp. 152-155, May 2017, doi: 10.5395/rde.2017.42.2.152.
[26] S. C. Gad and C. G. Rousseaux, "Use and Misuse of Statistics in the Design and Interpretation of Studies," in *Handbook of Toxicologic Pathology*, vol. 1, 2 ed., 2002, ch. 15, pp. 327-418.
[27] O. Smedby and M. Fredrikson, "Visual grading regression: analysing data from visual grading experiments with regression models," *Br J Radiol,* vol. 83, no. 993, pp. 767-75, Sep 2010, doi: 10.1259/bjr/35254923.

Annex

Statistical analysis

1. Interrater reliability: Cohen's Kappa

To test the interrater reliability, the kappa test (κ) is a frequently used statistical test. Traditionally, the interrater reliability was measured as percent agreement which is the number of agreement scores divided by the total number of scores. The Cohen's kappa take into account the uncertainty with which a rater gives an answer [22]. It is commonly used for assessing nominal variables [19].

The kappa coefficient can vary from -1 to +1 where 1 represents perfect agreement. 0 represents the degree of agreement due to the probability of chance or random agreement. Interpretation of the kappa is the same across multiple studies because this is a standardized value, as in other correlation tests. The kappa results are interpreted as follow [22]:

- ≤ 0: no agreement
- 0.01–0.20: none to slight
- 0.21–0.40: fair
- 0.41– 0.60: moderate
- 0.61–0.80: substantial
- 0.81–1.00: almost perfect agreement

Low negative kappa values (0 to -0.10) represents disagreement and indicates a problem. Great disagreement among raters is represented by a large negative kappa [22]. Kappa is computed according to the follow formula:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (1)$$

P(a) represents the observed agreement and P(e) the chance agreement. By cross-tabulating the rating of two raters, one can determine the degree of observed agreement. The chance agreement is provided by the marginal frequencies of the raters ratings. Cohen's kappa is mathematically only suitable for 2 observers. When 3 or more observers are included one can compute the arithmetic mean of the kappa's obtained between pairs. When working with categorical data with an ordinal structure, a weighted kappa is used. Here, the interrater-reliability estimates of the raters are assigned a different weight when their answers are closer or further apart. For example when one observer assigns 'Perfect' to an image and the other 'Inadequate' this will result in a lower IRR estimate than when one assesses 'Perfect' and the other 'Good' [19].

2. The chi-square test

The chi-square (χ^2) test is a nonparametric statistical analysing method to assess if two or more classifications of samples are independent or not [23]. For testing hypotheses on nominal variables, this test is very useful and provides information on the significance of the observed differences [24].

The chi-square test can be used when the data meets the following conditions [24]:

1. Variables can be nominal or ordinal.
2. The sample sizes of the study groups may be of equal or unequal size, unlike some parametric tests that must always have equally sized groups.
3. The original data were measured at an interval or ratio level, but violate one of the following assumptions of a parametric test, listed below:
    a. A distribution free statistic rather than a parametric statistic must be used when the data is skewed.
    b. The assumptions of equal variance or homoscedasticity of the data could not be met.
    c. The data is no longer ratio or interval because the continuous data were disintegrated into a small number of categories.

The outcome of the χ^2-test is a χ^2-value which enables us to make a comparison between the observed and expected frequencies. The formula is presented below (formula 1), where O is the observed value, E is the expected value, and S is the sum of all cells in the table. In order to define the significance level of the statistical test, one must calculate de degrees of freedom (df) and check this with the chi-square table. The degrees of freedom is calculated with formula 2. The χ^2-value corresponding with a certain degree of freedom determines the probability level (p) [23, 24].

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{2}$$

$$\text{df} = (\text{Number of rows} - 1) \times (\text{Number of columns} - 1) \tag{3}$$

3. Fisher's exact test

Whereas the chi-square test only says something about the independent or associated relationship between different variable, the Fisher exact test can do a pairwise comparison between the variable. The Fisher exact test can be used as a *post-hoc* test [25]. The sample size used for the test should be small but can actually be calculated for any size of sample. The data is discontinuous categorical data and should be independent of each other. The exact test produces a probability (p) that determines whether the groups differ significantly or not. It is calculated with the following formula [26]:

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!} \quad (4)$$

To make the calculations, the data should be in a 2x2 contingency table (table 2). By applying hypergeometric distribution of the numbers in the cells of the table, the Fisher's exact test assesses the null hypothesis of independence. When p<.05 the null hypothesis can be rejected and so one can conclude that there is evidence of a statistical significant difference between groups giving a positive answer [25, 26].

Table 22: 2x2 contingency table

Group	"Positive"	"Negative"	Total
Group 1	A	B	A+B
Group 2	C	D	C+D
Totals	A+C	B+D	A+B+C+D = N

4. Ordinal logistic regression

When one wants to analyse a non-linear relationship between two variables, ordinal logistic regression can be applied on the data. A non-linear relationship accounts for the fact that a change in one variable (e.g., compression force) does not produce a directly proportional change in the other (e.g., PGMI score) [26]. The ratio between the probability that an event occurs and the probability of not occurring is called the *odds*. Transforming the probability (formula 5) with the logistic function results in a linear equation (formula 6), with which one continuous independent variable takes the form noted in formula 7.

$$logit\ (p) = \log\left(\frac{p}{1-p}\right) \quad (5)$$

$$logit\ (p) = ax + b \quad (6)$$

$$p = \frac{1}{1+\exp(ax+b)} \quad (7)$$

When the independent variables are categorical they should be represented by a term that takes a separate value for each category. When there are multiple independent variables, a linear combination of the independent variables is needed. This is represented by formula 8 where 'z' is a weighted sum of independent numerical variables [27].

$$p = \frac{1}{1+\exp(-z)} \quad (8)$$