Research article

Development and content validity evaluation of a candidate instrument to assess image quality in digital mammography: A mixed-method study

Joana Boita a,b,1, Anetta Bolejko c,1, Sophia Zackrisson c, Matthew G. Wallis d, Debra M. Ikeda e, Chantal Van Ongeval f,1, Ruben E. van Engen b, Alistair Mackenzie g, Anders Tingberg b, Hilde Bosmans b,1, Ruud Pijnappel b,1, Ioannis Sechopoulos a,b, Mireille Broeders b,1, Mireille Broeders a

A R T I C L E  I N F O

Keywords:
Digital mammography
Image quality evaluation
Content validity evaluation
Directed content analysis
Content validity index

A B S T R A C T

Purpose: To develop a candidate instrument to assess image quality in digital mammography, by identifying clinically relevant features in images that are affected by lower image quality.

Methods: Interviews with fifteen expert breast radiologists from five countries were conducted and analysed by using adapted directed content analysis. During these interviews, 45 mammographic cases, containing 44 lesions (30 cancers, 14 benign findings), and 5 normal cases, were shown with varying image quality. The interview findings were used to develop tentative items, which were evaluated in terms of wording, understandability, and ambiguity with expert breast radiologists. The relevance of the tentative items was evaluated using the content validity index (CVI) and modified kappa index (k*).

Results: Twelve content areas, representing the content of image quality in digital mammography, emerged from the interviews and were converted into 29 tentative items. Fourteen of these items demonstrated excellent CVI ≥ 0.78 (k* > 0.74), one showed good CVI < 0.78 (0.60 ≤ k* ≤ 0.74), while fourteen were of fair or poor CVI < 0.78 (k* < 0.59). In total, nine items were deleted and five were revised or combined resulting in 18 items.

Conclusions: By following a mixed-method methodology, a candidate instrument was developed that may be used to characterise the clinically-relevant impact that image quality variations can have on digital mammography.

1. Introduction

Optimal performance of digital mammography (DM) screening is crucial for detecting breast cancer at an early stage. For this, the performance of DM systems, including their corresponding image processing, needs to be tested and optimised, taking into account the issues that
actually affect image interpretation and breast cancer detection. However, currently used image quality metrics do not account for these issues [1–6].

Before new or modified DM systems are introduced into screening, physical and clinical assessments are undertaken in order to verify if the systems fulfil the stated performance and clinical requirements. Examples of such an evaluation procedure are type testing described in the Supplement to the European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis [7] and FDA-required Equipment Evaluation testing for DM systems [8]. In type testing, two complementary evaluation phases are performed: a physics-based assessment, and a clinical assessment, derived from objective and subjective evaluations, respectively.

The former is used to assess and characterise the performance and technical capabilities of DM systems using technical measurements on the raw, “for processing”, phantom images.

The second phase consists of assessing the image quality of the processed, “for presentation”, DM images from a clinical point of view. In this phase, human observers identify any issues that may arise as a consequence of inadequate image processing algorithms by rating the representation of normal anatomical structures in the images. Their subjective opinions on image quality are quantified as follows: observers are asked to score a set of criteria regarding physical parameters (resolution, noise, and contrast) and relevant normal anatomical structures reproduced in the digital mammogram. However, specific features from breast lesions are not included. The overarching assumption in the design of this clinical evaluation is that the probability of detecting a lesion is correlated to the representation and visibility of normal anatomical structures [9,10]. In addition to the set of criteria used in the clinical evaluation of type testing, there are two other sets of image quality criteria, one introduced by Hemdal et al. [11] and the other by Van Ongeval [4,12], that cover similar content: physical parameters, relevant anatomical structures, and non-lesion features. These two sets were developed on the basis of the image quality criteria recommended by the previous European Guidelines on Quality Criteria for Diagnostic Radiographic Images for conventional mammography [13], to assess image quality in digital mammography. They were used in studies to investigate the impact of dose reduction on the quality of DM images [11,12] and to compare different DM systems and different image processing algorithms in digital mammography [12]. In addition, another set of criteria was proposed for assessment of mammographic image quality, in an effort to update and adapt the criteria used in the National Health Service Breast Screening Programme (NHSBSP) guidelines to the digital era [14–16].

The clinical assessment of image quality in type testing can be considered an extension of Visual Grading Analysis (VGA), which is a simple and intuitive method to assess image quality, consisting of visually grading the visibility and reproduction of clinically important anatomical structures [17–19]. The low effort needed to perform a VGA test makes it suitable for use in type testing. However, VGA tests are subjective and have a tendency to result in a “beauty contest”, i.e., issues not relevant to detection or characterisation of lesions can lead to low scores because there is a tendency for an observer to evaluate an image in terms of aesthetics or force of habit rather than clinical purpose. The same issues arise when directly comparing two different images in terms of image quality that are actually of equivalent underlying diagnostic performance as there is a tendency for the observer to prefer one image over the other for the same aesthetic or familiarity reasons [19,20]. Finally, this type of assessment does not predict clinical performance; i.e., it is not clear if passing this subjective evaluation guarantees that a system will provide images suitable for the clinical task [21].

This clinical assessment is an essential step in image quality evaluation because it allows for testing of conditions equivalent to those when the system is in use. It also allows for testing of how DM images would perform in clinical conditions, which is the goal of optimising parameters tested during the physics-based measurements. Consequently, it is essential to have a valid set of image quality criteria that accurately correlate with actual lesion detection and diagnosis performance in digital mammography.

Therefore, the aim of this study was to develop a candidate instrument to assess image quality in digital mammography from the perspective of the radiologists, by identifying clinically relevant features in DM images that are affected by lower image quality.

2. Materials and methods

2.1. Study overview

In this mixed-method study, qualitative and quantitative methods were used for development and content validity evaluation of a candidate instrument aimed to assess image quality in digital mammography (the phenomenon). The instrument is denoted as being a candidate because it is not yet fully validated in terms of construct validity and reliability [22]. To develop the instrument and evaluate its content validity, face-to-face interviews with expert breast screening radiologists were conducted and analysed according to the adapted directed content analysis [23]. During the interviews, digital mammograms at varying levels of image quality were shown to the radiologists in order to investigate the construct of image quality in digital mammography from the perspective of the radiologists, that is, to identify clinically relevant anatomical features when interpreting mammograms, and to investigate how these features are described. In addition, detailed explanations and descriptions of the impact of lower image quality on those features were also investigated. Triangulation, in terms of different researchers performing the analysis, was carried out as a trustworthiness check of the findings [24]. The instrument was built by creating tentative items based on the interview findings. Content validity was quantified and also evaluated by means of CVI to assess the relevance of each tentative item in the instrument [22,25,26]. In addition, wording, understandability, and ambiguity of the tentative items were also investigated [22,27,28]. To complete the development of the candidate instrument, the rating scale was selected. Fig. 1 presents a brief overview of the development and content validity evaluation of the resultant items in the candidate instrument. Details on the content validity evaluation can be found in the online supplement.

2.2. Preparation of the digital mammography images

2.2.1. Digital mammography images selection

A total of 45 DM cases with breast density varying from fatty to dense, with and without lesions, acquired using Lorad Selenia (Hologic, Inc., Bedford, MA) DM systems, were selected and retrieved, under licence, from the OMI-DB anonymised database of mammograms, which is part of the OPTIMAM project [29]. Each case consisted of two views (cranio-caudal and mediolateral oblique) of each breast. This set contained a total of 44 lesions (30 cancers and 14 benign findings) and included 5 normal cases. The positive cases contained different types of lesions, in order to include most of the types of lesions found at screening. All cases had pathological (in case of the lesion-containing cases) or follow-up confirmation. The cases with lesions were reviewed and annotated by an expert breast radiologist.

2.2.2. Modification of the quality of the mammograms

Algorithms were previously developed [30–32] and validated [33] to modify the quality of mammograms by simulating five previously observed image quality issues in DM systems submitted for type testing. Specifically, resolution was reduced; contrast was increased; contrast was decreased; texture of the structures was modified by increasing the correlation of the pixel values in the image; and noise was added to simulate lower-dose acquisitions.

The quality of all the mammographic cases was reduced by simulating one type of image quality issue per case. For each of the five
simulated image quality issues, six images with increasing levels of degradation were generated from the reference image. The reference image, i.e., the image without any change in quality, was determined to be adequate for interpretation, but not necessarily perfect, as is the case with the majority of acquired mammographic images. Out of the four views in each reference case, the image to be degraded was chosen to be the view in which the lesion features, when present, were most visible. The levels of degradation were selected according to what is clinically relevant and realistic. In total, seven images (reference image in addition to the six degraded images) for each DM case and each image quality issue were prepared to be shown to the radiologists during the interviews.

2.3. Radiologists’ interview

The interviews were conducted in English, by the same interviewer, with fifteen expert breast screening radiologists from five countries (UK, USA, Sweden, Belgium, and the Netherlands) with a median of 20 years (range: 5–46 years) of experience in reading mammograms. The criteria for radiologist expertise were having at least 5 years of experience in breast imaging and being involved in regular reading of screening mammography. The number of radiologists was selected based on existing literature [34]. All radiologists were interviewed individually. The purpose was to investigate the construct of image quality in digital mammography from the perspective of the radiologists. That was done by identifying the features in breast lesions and normal breast tissue that are relevant to the radiologists when interpreting a mammogram, and how those features are affected by lower image quality. Also, how the radiologists described the features in detail and explained the impact of lower image quality in those features was investigated.

Three test interviews were first conducted to verify that the interview setup was sufficient to obtain the information intended and so that the previously-inexperienced interviewer could become familiar and comfortable with the interview workflow. The outcomes of these test interviews were reviewed with the study team, which includes a researcher highly experienced in qualitative content analysis, and, especially, the potential pitfalls when interviewing were discussed in order to gain awareness of how to handle them. Based on this, the number of cases to include, the duration of the interview, and the optimal way to display the images were defined. In addition, the test interviews were used to develop and improve a written interview guide that was used to conduct all the subsequent fifteen interviews analysed in the study. These test interviews were conducted with one radiologist from within and two from outside the group of fifteen radiologists that participated in the actual interviews analysed.

Each of the fifteen interviews took approximately four hours, divided in two sessions of approximately two hours each. Short breaks (~10 min) where possible at the request of the radiologists throughout the session, while one mandatory ~1-h break was taken between the two sessions. For three interviews, the two sessions were performed in
two different days. They were performed in a quiet room with suitable ambient light conditions for image reading, and audio and video recorded. During each interview, the DM cases were viewed on either 5- or 12-megapixel calibrated liquid crystal diagnostic mammography displays (Barco, Kortrijk, Belgium). The images were displayed according to the DICOM standard for presentations using ViewDEX [35], a software specifically developed for observer studies. The software allowed the radiologists to move forward and backward between cases, and to zoom, pan, and scroll over the images of each case.

For each case, first, the reference images were shown on one program window, and the location of the lesion was indicated by the interviewer. The radiologist was asked to review and analyse the full case by identifying and giving extensive explanations about the relevant features and their visibility considering the image quality of the reference images. Second, the radiologist was presented with the seven mammograms shown in decreasing image quality order in another program window displayed next to the first one. The first image of each set of images was the reference image. The radiologist was asked to give explanations about the effect of lowering image quality on the visibility of the structures and features reported in the first phase, and to identify the image they felt was still acceptable for image interpretation. Additionally, he/she was asked to explain in detail why he/she was still confident with the selected image to look for other lesions. When looking at the first image the radiologist felt was no longer acceptable for image interpretation, he/she was asked if any of those important features reported earlier were missed in that image. They were also asked to explain the reasons that made that image unacceptable.

During both phases, probing questions were constantly used to help the radiologists to talk extensively about what they were observing, and to explain it with different wording and with details.

2.4. Analysis of the interviews and development of the tentative items

The interviews were transcribed verbatim (Mijntranscript, The Netherlands). The transcripts were verified, comparing each of them to the respective audio. As described in the online supplement, directed content analysis methodology [23] was adapted to analyse the transcripts by merging it with content validity evaluation. This allowed for interpretation of the text data corresponding to the detailed explanations and descriptions given by the radiologists about the construct of image quality in digital mammography.

The transcript analysis began with the process of familiarising oneself with the data, where the transcripts were first read several times to gain a sense of the text as a whole, while keeping the goal of the study in mind [36,37]. While reading the transcript several times, the text was divided into meaning units, corresponding to the radiologist’s descriptions of breast tissue and lesion features, and explanations of the impact of each type of degradation on those features. The meaning units were condensed while still keeping the content regarding a particular feature. During coding, the codes were developed, allowing to obtain an understanding of the construct of image quality (Table 1S in the online supplement). The codes were organized into clusters, that related to each other through their content, and then combined into content areas of image quality in digital mammography. The transcripts were checked and rechecked several times, and the codes were compared in order to verify that all the information related to the construct of image quality was included. Incorporating the wording used by radiologists, the tentative items where then created reflecting the content found in the clusters of codes from one or more content areas [36,37].

2.5. Triangulation in the interview analysis and in the development process

One researcher analysed all the interviews. However, the four interviews that seemed most informative were also analysed independently by another researcher experienced in qualitative content analysis, and the tentative results obtained by the two researchers were compared and discussed. In addition, an experienced breast radiologist, who participated in the interviews, independently performed the analysis of one interview conducted with another radiologist, and the understanding of the findings was discussed. This discussion confirmed the interpretation of the findings obtained by the two researchers and their respective analysis. Doing so, triangulation in terms of peer debriefing was performed [24,37]. Then, the results from all the interviews were discussed among the entire interdisciplinary team, which is another form of peer debriefing. These discussions also allowed to look at the findings from different perspectives and share different understandings of the interview findings [24,37].

As explained in the online supplement, besides the qualitative data, the quantitative data obtained from the CVI analysis was included in the development process, which is another type of triangulation to strengthen the trustworthiness of the study results [24,36].

2.6. Content validity index, verification, and revision of the tentative items

The tentative items of the instrument were reviewed in terms of wording, understandability, and ambiguity by four expert breast radiologists, including a native English speaker [28]. The item relevance was investigated by using the CVI [25,26]. Also, the understandability was indirectly tested: in case an item was rated as not relevant by some experts, but it was consistently referred to during the interviews, the reason it was rated as not relevant might have been because it was not understood [25,26].

Eight of the fifteen expert breast radiologists, that had participated in the interviews and that were clearly informed about the goal of the instrument, evaluated and rated the relevance of each tentative item on an online survey platform (SurveyMonkey Inc., San Mateo, California) using a rating scale from 1 to 4 (1—not relevant, 2—somewhat relevant, 3—quite relevant, 4—highly relevant) [25]. The number of experts was selected considering that, the higher the number of experts, the lower the total agreement among them, which increases the agreement by chance [25].

I-CVI (Item-Content Validity Index) was calculated as the number of experts giving a rating of either 3 or 4, divided by the number of experts, that is, the proportion in agreement that the item is relevant. I-CVI is also an index of inter-rater agreement on relevance of each item that describes the construct of the phenomenon. It was complemented by the modified kappa index (k*), because it provides information about the degree of agreement on relevance beyond chance [26]. This index can be estimated through the following equation:

\[
k* = \frac{(I-CVI - p_c)/(1 - p_c)}{p_c} = [N(A - 1)!(N - A)!] \times 0.5^N
\]

where pc is the probability of a chance occurrence:

\[pc = [N(N - 1)!/A!(N - A)!] \times 0.5^N\]

where N is the total number of experts and A is the number of experts agreeing on good relevance, that is, number of experts giving a rating of 3 or 4. A minimum I-CVI value of 0.78 or higher indicates that the respective item is acceptable. Regarding the modified kappa index, an item is considered excellent if k* value is higher than 0.74; good if k* value is between 0.6 and 0.74; fair if k* value is between 0.40 and 0.59; and poor if k* value is lower than 0.40 [25,26].

Lastly, the qualitative and quantitative results were combined to revise and to include or reject a certain tentative item. A tentative item was included if the I-CVI and k* scores were equal or higher than 0.78 and 0.60, respectively, without the need for rephrasing. When the calculated values of I-CVI and k* were lower than 0.78 and 0.59, respectively, the tentative item was verified, in an attempt to understand the reasons for the low score. With this understanding, it was possible to decide if the items should still be included but revised, or deleted. For
example, an item could still be included by confirming if it was reflecting an important feature reported during the interviews or if three experts, who participated in the CVI evaluation and in the final revision of the items, agreed on still keeping it. However, this item might need to be rephrased and modified, because the lower scores could mean lower understandability or higher ambiguity [39].

2.7. Development of the rating scale

The development of the instrument, together with the creation of the items, included the selection of a rating scale. The choice of rating scale was directly related to the goal and the type of instrument, and based on the following assumptions: (1) the responses are on a discrete scale, since at the end a pass/fail score should be calculated; (2) neutral responses are not included, since the instrument will be used to approve or reject a mammography system submitted for type testing; (3) the use of scales with 7–10 scores results in a smaller loss in reliability than when using less than 7 scores; and (4) people find large scales difficult to use, and there is good evidence that, for a wide variety of tasks, people cannot discriminate well beyond seven levels, and do not select the most extreme scores [22]. Considering these four assumptions, 8 was the most suitable number of scores to be included in the scale, which, therefore, ranges from 1 to 8 (Completely disagree to completely agree), including the option NA (Non-applicable) for a case where an item cannot be answered (e.g. criteria about lesion in non-lesion cases). This type of scale is considered a bipolar scale, where one extreme score reflects strong accordance of an idea, and the other extreme category reflects strong accordance to its opposite [22].

3. Results

3.1. Findings from the interview analysis

Twelve content areas emerged from the interviews, as listed in Table 1, along with their respective condensed meaning units, codes, and corresponding tentative items that were developed on the basis of the interview findings are shown in Table 2.

3.2. Development and verification of the tentative items

In total, 29 tentative items (Table 3) were created from the content in the content areas, as previously showed in Table 2, and verified in terms of wording, understandability, and ambiguity.

If a code reflected a feature that is not normally affected by lower image quality or that could be covered by another tentative item, then that specific code was not converted into a tentative item. For instance, as seen in Table 1, in the content area “Direction”, the code “Direction of the calcifications” is taken into account when radiologists interpret mammograms. However, it does not reflect a feature that is directly affected by lower image quality and it can also be covered by the code “Distribution of the calcifications” in the content area “Distribution”. Additionally, the tentative items that emerged from the codes belonging to the content areas “Effects on the surroundings”, and “Surrounding Structures” were combined and resulted in tentative items about visibility of normal structures, and artifacts in the image. Two content areas “Location” and “Lesion vs Surrounding Structures” were removed because either the included meaning units could be placed in other content areas or the resultant tentative items were already covered by other content areas. Additionally, personal- and beauty-based classifiers and descriptors, such as “well/good visible”, “I can evaluate”, and “How confident are you (...)?” were avoided in phrasing the tentative items. All the tentative items were positively framed to represent the preferred scenario of the statement in each item. In the end, the tentative items were divided into three groups: calcifications, soft tissue lesions, and normal tissue, since the obtained features, and respective descriptions

| Table 1 | Summary of the content of each content area expressed by the codes. |
|---|---|
| Codes | Content Areas |
| Different densities among calcifications | Density |
| Differentiation between glandular tissue and lesion | Differentiation |
| Distribution of the calcifications | Distribution |
| Distribution of the breast tissue | Distribution |
| Gray-scale of the glandular tissue | Number |
| Number of lesions | Number |
| Different morphologies among calcifications | Morphology |
| Visibility of lesion morphology | Size |
| Different sizes among lesions | Size |
| Visibility of lesion margins | Margins |
| Visibility of irregular lines | Lines |
| Visibility of ducts (in terms of sharpness and in comparison to surroundings) | Lines |
| Visibility of vessels (in terms of sharpness and in comparison to the surroundings) | Lines |
| Visibility of Cooper’s ligaments (in terms of sharpness and in comparison to surroundings) | Lines |
| Direction of the calcifications | Direction |
| Direction of irregular lines | Direction |
| Direction of breast tissue pattern | Direction |
| Direction of ducts | Direction |
| Direction of vessels | Direction |
| Direction of Cooper’s ligaments | Direction |
| Differentiation between breast tissue and lesion | Lesion vs Surrounding Structures |
| Differences in structure, architecture, texture, and arrangement of the breast tissue | Lesion vs Surrounding Structures |
| Distinction of adipose and glandular tissue | Location |
| Visibility of the vessels | Location |
| Visibility of the ducts | Location |
| Visibility of normal lines | Location |
| Visibility of nipple | Location |
| Visibility of skin line | Surrounded Structures |
| Visibility of pectoralis muscle | Surrounded Structures |
| Distinction of adipose and glandular tissue | Effects of the pectoralis muscle |
| Distortion of the normal lines | Effects on the nipple |
| Effects of the pectoralis muscle | Effects on the nipple |
| Retraction of the skin line | Effects on the nipple |
| Distortion of the architecture and texture of the breast tissue | Effects on the nipple |

and explanations, corresponding to each group are different.

3.3. Content validity index of the tentative items

Table 3 shows the results from the CVI evaluation of the 29 tentative items performed with 8 expert breast radiologists. 14 tentative items demonstrated excellent content validity (I-CVI ≥ 0.78, k* > 0.74), and therefore were considered relevant. One tentative item showed good content validity (I-CVI < 0.78, 0.60 ≤ k* ≤ 0.74), and 14 tentative items were considered fair or poor (I-CVI < 0.78, k* ≤ 0.59).

3.4. Revision of the tentative items and development of the candidate instrument

Tentative items 7, 12, 14, 18, 19, 21, 22, 24, and 25 were deleted, because they were not considered relevant to be included, or because they were redundant. Tentative items 17, 20, and 23 were merged into one item. Tentative item 27 was rephrased in order to clarify the codes they reflected. Although tentative item 26 was only considered a fair item, it was still kept without any change because it was an important feature reported in the interviews and experts agreed to keep it in order to include items about the important structures of the breast. Table 4 shows the revised 18 items that made up the resulting instrument.

4. Discussion

An image presents good quality if it allows for the distinction and representation of clinically relevant structures and features [20].
Therefore, it is important to understand, beforehand, what kind of structures and features related to breast lesions and breast tissue are important to radiologists when interpreting a mammogram. In this study, those structures and features were identified, based on detailed descriptions and explanations provided by expert breast radiologists from several countries. The interview findings provided an understanding of those features, such as ‘distribution’, ‘density’, ‘margins’, ‘morphology’, etc., that should be, therefore, considered when assessing image quality in digital mammography (shown in Table 1). This information was used to develop an instrument composed of 18 items. The items were divided into three groups: calcifications, soft tissue lesions, and normal tissue, since the obtained features corresponding to each group are different. It is important to note that the items are not directly about physical parameters. However, each item is about a feature that allows to evaluate image acquisition-related characteristics and post-acquisition image processing-related indirectly. For instance, when evaluating the texture of the breast tissue (Table 4, item 1) or the distinction between adipose and glandular tissues (Table 4, item 3), it is intended to evaluate, respectively, the presence of noise, and the contrast in the image. A drawback of visual grading studies is related to the preference for one image over another resembling a ‘beauty contest’ [19,20]. In the development of the items, there was an attempt to reduce that effect, by avoiding personal- and beauty-based classifiers and descriptors.

The study findings are, for the most part, in agreement with previous knowledge about physical parameters and their effects on clinically relevant anatomical structures, as described in the criteria to evaluate image quality previously developed by Hemdal et al. [11] (set1), by Van Ongeval et al. [4,12] (set 2) and its adaptation included in the clinical evaluation section of the European type testing [7] (set 3), and the set proposed [14][15] for adaptation of the UK NHSBSP guidelines [16] (set 4) shown in Table 4. However, there are differences between the candidate instrument developed in this study and the previous sets in how the content of that previous knowledge is covered. The candidate instrument developed in this study includes items about specific lesion features, discriminating between calcifications and soft tissue lesions. This is a relevant addition, because the previous assumption, that visualisation of normal structures correlates with that of lesion structures has been shown to not be true [6,40]. Additionally, in the candidate instrument, items about physical parameters were not included, whereas sets 2 and 3 did include a “Physical characteristics/parameters of an image” section. This was taken into account when developing the candidate instrument, because a VQA test, by definition, evaluates the reproduction of anatomical structures in clinical images, which are influenced by physical parameters, but these are not asked about directly. This is an advantage since the interpretation of the meaning and impact of each physical parameter may be different across clinical observers. Also, the items are, in general, about specific features and not about overall evaluations, as seen in the previous criteria such as “How satisfied are you with the representation of the image?” (Table 4, set 3, item 7). In contrast to set 4, the items in the candidate instrument were phrased avoiding personal- and beauty-based classifiers and descriptors, which could create ambiguity. For example, the answer to an item like “I can see the greys of the glandular tissue.” may not reflect the assessment of differences between adipose and glandular tissues, but the ability of the observer to assess the differences between the two tissues. Finally, this instrument does not include any item about breast positioning, whereas sets 1 and 4 included positioning-related items. In the development of the candidate instrument, those types of criteria were not included because they do not relate to the performance of the digital mammography systems per se. Therefore, those types of issues should be addressed in other evaluation procedures [41,42].

One of the strengths of the study is the mixed-method research approach used in the development and content validity evaluation of the instrument [38]. The combination of qualitative data from the interviews with the quantitative data from the CVI allowed to take decisions on keeping or removing a certain item. For instance, if an item of poor CVI reflected an important feature reported during the interviews, it was kept in the instrument. The use of two different methodologies is a type of triangulation. Triangulation was an important process in the analysis of the interview data that allowed for verifying the veracity of the obtained data with other experts, especially when using different approaches and exploring the phenomenon from different perspectives [24,37,38]. Following fifteen interviews, saturation of the data was observed, meaning that no new data could be obtained beyond a certain point [43]. This indicates that the number of participating radiologists was adequate [43]. Another strength of the study is the diversity of the group of radiologists, in terms of experience, country of origin and practice, and breast cancer screening program characteristics. This contributed to obtaining heterogeneous data, i.e., different descriptions and interpretations of the same features [36]. Finally, the diversity of digital mammography cases and types of degradation included in the study allowed for replicating different conditions observed in clinical and system-evaluation setting.

This study also has some limitations: the interviews were conducted in English and most of the participants were English-speaking and the group of radiologists who conducted these interviews were not native English speakers. Although it was assumed that the understanding of the construct of image quality is independent of language and that all participants had good English language skills, the capabilities and confidence to express and explain an idea in more detail may be affected when it is done in a language that is different from their native language [22]. However, the same findings emerged across radiologists of different native language. In addition, the ability of the developed instrument to actually predict

### Table 2
Examples three content areas and respective condensed meaning units, codes, and tentative items.

| Condensed Meaning Units | Codes | Content Areas | Tentative Items |
|-------------------------|-------|---------------|-----------------|
| 'They (calcifications) have a linear distribution.' | Distribution of calcifications | The distribution of the calcifications is visible. |
| ‘(...) the nice, even black and white distribution of the tissue.' | Distribution of breast tissue | The texture of glandular and adipose tissues is depicted appropriately throughout the image. |
| 'There is a reasonably regular pattern on glands and ducts and fat.' | Distribution of breast tissue | Distribution |
| 'All the lines are going to the right places; the ducts and the fat radiate from behind the nipple.' | Gray-scale of the glandular tissue | The grey-scale depiction of glandular tissue is appropriate. |
| 'calcifications) They have different shapes from each other.' | Different morphologies among calcifications | Morphology |
| 'Punctate calcifications.' | Visibility of lesion morphology | The morphology of the calcifications is visible. |
| 'Well circumscribed mass.' | 'Classical tent sign.' | The morphology of the soft tissue lesions is visible. |
| 'Irregular border.' | 'The edge is a little bit darker than the centre of the breast.' | The margins of the soft tissue lesions are depicted sharply. |
clinical performance and the relationship of the items to the instrument has not been tested yet. The clinical relevance of this instrument, i.e., its correlation with detection and diagnosis performance in digital mammography is currently being tested, using receiver operating characteristic-based methods. Also, whether the items in the instrument assess the same construct, if they provide equal amount of information about the construct of image quality, and how well the instrument is measuring what it should is being investigated.

5. Conclusions

In this mixed-method study, relevant structures from breast tissues and breast lesions, that are affected by lower image quality, were identified and used to develop a candidate instrument. Subject to validation, this instrument will likely have utility in the assessment of image quality in digital mammography and to characterise the clinically-relevant impact that image quality variations can have on digital mammography. This study also illustrated the value of qualitative methodology in image quality evaluation studies.

Statement of ethics

The authors have no ethical conflicts to disclose.

Funding source

Alistair Mackenzie was funded as part of the OPTIMAM2 project and is supported by Cancer Research UK (grant number: C30682/A17321). This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Table 4: Image quality criteria previously developed and the 18 items from the candidate instrument developed in this study.

| Image quality criteria related to positioning | Image quality criteria developed by Hemdal et al. [11] (set 1) | Image quality criteria developed by Van Ongeval et al. [4,12] (set 2) | Image quality criteria used in European type testing [7] (set 3) | Image quality criteria suggested by Mercer et al. [14] and developed by Moran [15] (set 4) | Candidate instrument developed in this study |
|---------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------|------------------------------------------------|
| 1.1 Pectoral muscle at correct angle        | 1. Visualisation of the skin line                             | 1. Is the skin line visualised?                                 | 1. I am confident that the skin surface is visualized on the cranio-caudal projections. | Normal Tissue                      |
| 1.2 Infra-mammary angle visualised          | 2. Visualisation of the vascular structures through the dense parenchyma | 2. Are the vascular structures visible through the dense parenchyma? | 2. I am certain there is no movement on the images. | 1. The texture of glandular and adipose tissues is depicted appropriately throughout the image. | 2. The greyscale depiction of glandular tissue is appropriate. |
| 1.3 Nipple in profile, clear of overlying breast tissue and/or indicated by marker | 3. Sharp visualisation of vascular and fibrous structures and pectoral muscle | 3. Is the pectoral muscle visualized sharply? | 3. I am not confident to evaluate the lower half of the breast on the left medio-lateral oblique projection. | 3. The distinction between adipose and glandular tissues is appropriate. | 4. The normal structures, such as the Cooper’s ligaments, ducts, and vessels, if present, are depicted sharply. 5. The greyscale depiction of the pectoralis muscle, if present, is appropriate. |
| 1.4 No skinfolds seen                      | 4. Sharp reproduction of skin structures along the pectoral muscle | 4. Are the Cooper’s ligaments and vascular structures in the subcutaneous and pre-pectoral area visualized well? | 4. I am confident that the whole breast is imaged on the medio-lateral projections. | 6. Soft Tissue Lesions (mass and architectural distortion) |
| Image quality criteria related to detector performance, exposure parameters and patient movement | 5. Visualisation of the Cooper’s ligaments and vascular structures in the white areas | 5. Are the microcalcifications well visualized and well outlined? | 5. I am confident that the skin surface is visualized on the medio-lateral projections. | 7. The nipple is visible. | 8. There are no artefacts in the image. |
| 2.1 Visually sharp/clear reproduction of glandular tissue | 6. Visualisation of the Cooper’s ligaments and vascular structures in the dark areas | 6. Is the contrast in the dark areas sufficient (e.g. no saturation of intensity of signals, no fully dark regions)? | 6. I am not able to visualise vessels through the densest glandular tissue on the cranio-caudal projections. | 9. The morphology of the calcifications is visible. | 10. The morphology of the calcifications is visible. |
| 2.2 Visually sharp/clear reproduction of fibrous strands in fat tissue | 7. In the white areas, no disturbing noise (pseudomicrocalcifications due to noise) | 7. Is the contrast in the white areas sufficient (e.g. no fully white regions)? | 7. I can differentiate between fatty and glandular tissue on the medio-lateral projections. | 11. Calcifications are shown with a different density compared to that of the surrounding tissue. | 11. Calcifications are shown with a different density compared to that of the surrounding tissue. |
| 2.3 Visually sharp/clear reproduction of vascular structures in fat tissue | 8. In the dark areas, no disturbing noise (pseudomicrocalcifications due to noise) | 8. Is the glandular tissue sufficiently white? | 8. The morphology of the calcifications is visible. | 12. Within a cluster of calcifications, the individual calcifications are distinguishable from each other. | 12. Within a cluster of calcifications, the individual calcifications are distinguishable from each other. |
| 2.4 Visually sharp/clear reproduction of pectoral muscle margin | 9. Enough contrast in the white areas | 9. Is the background sufficiently dark? | 9. The distribution of the calcifications is visible. | 13. Within a cluster of calcifications, differences in density among calcifications are visible. | 13. Within a cluster of calcifications, differences in density among calcifications are visible. |
| 2.5 Visually sharp/clear reproduction of calcifications, when present | 10. Enough contrast in the dark areas | 10. Do all images appear in the same way? (if not, please comment) | 10. The morphology of the calcifications is visible. | 14. Within a cluster of calcifications, differences in morphometry among calcifications are visible. | 14. Within a cluster of calcifications, differences in morphometry among calcifications are visible. |
| 2.6 Acceptable noise level in the reproduction of the pectoral muscle | 11. The glandular tissue is white | 11. Is there disturbing noise in the dark areas? | 11. The morphology of the calcifications is visible. | Soft Tissue Lesions (mass and architectural distortion) | 15. Soft tissue lesions are shown with a different density compared to that of the surrounding tissue. |
| Overview judgement related to either screening or diagnostic procedures | 12. The background is dark | 12. Is there disturbing noise in the white areas? | 12. The morphology of the soft tissue lesions is visible. | 16. The morphology of the soft tissue lesions is visible. | (continued on next page) |
Table 4 (continued)

| Image quality criteria developed by Hemdal et al [11] (set 1) | Image quality criteria developed by Van Ongeval et al [4,12] (set 2) | Image quality criteria used in European type testing [7] (set 3) | Image quality criteria suggested by Mercer et al [14] and developed by Moran [15] (set 4) | Candidate instrument developed in this study |
|---|---|---|---|---|
| How confident are you with the representation of the microcalcifications? | 5. How satisfied are you with the representation of microcalcifications? | 20. The glandular breast tissue is not sufficiently penetrated to allow thorough evaluation of the breast in the cranio-caudal projections. | 17. The margins of the soft tissue lesions are depicted sharply. |
| How confident are you with the representation of opacities? | 6. How satisfied are you with the representation of opacities? | 21. I am confident that the pectoral muscle angle on the right medio-lateral oblique projection will not obscure breast tissue. | 18. Irregular/abnormal lines (that are disturbing the surrounding tissue) are depicted sharply. |
| How confident are you with the representation of the image? | 7. How satisfied are you with the representation of the image? | 22. I am confident that the whole breast is imaged on the cranio-caudal projections. | |
| | | 23. The glandular breast tissue is sufficiently compressed to allow for thorough evaluation in the cranio-caudal projections. | |
| | | 24. I am confident the quality of this set of images is sufficient to enable diagnosis through image interpretation. | |
| | | 25. The BI-RADS classification of the breast composition would be classified as: A, B, C, and D. | |

CRediT authorship contribution statement

Joana Boita: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. Anetta Bolejko: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. Sophia Zakrison: Conceptualization, Methodology, Writing - review & editing. Matthew G. Wallis: Conceptualization, Methodology, Writing - review & editing. Anders Tingberg: Conceptualization, Methodology, Writing - review & editing. Chantal Van Ongeval: Conceptualization, Methodology, Writing - review & editing. Ruben E. van Engen: Conceptualization, Methodology, Writing - review & editing. Alistair Mackenzie: Conceptualization, Methodology, Writing - review & editing. Hilde Bosmans: Conceptualization, Methodology, Writing - review & editing. Ruud Pijnappel: Conceptualization, Methodology, Writing - review & editing. Ioannis Sechopoulos: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing. Mireille Broeders: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing.

Declaration of Competing Interest

Some of the authors of this manuscript declare relationships with companies: Ioannis Sechopoulos: research agreements, Siemens Healthcare, Canon Medical Systems, ScreenPoint Medical, Sectra Benelux, and Volpara Health Technologies, and speaker agreements, Siemens Healthcare; Mireille Broeders: speaker agreements, Siemens Healthcare and Hologic; Sophia Zakrison Speaker agreements Siemens Healthcare, research agreement ScreenPoint Medical; Anders Tingberg: Research agreements, Siemens Healthcare; Matthew Wallis’ institution has received grants from Philips; Chantal Van Ongeval: speaker agreements Siemens Healthcare; Hilde Bosmans: research agreements Siemens Healthcare and GE Healthcare; Ruud Pijnappel: speaker agreement Hologic; and Debra Ikeda: consultant to Hologic. Alistair Mackenzie was funded as part of the OPTIMAM2 project and is supported by Cancer Research UK (grant, number: C30682/A17321). For the remaining authors none were declared. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

For the VISUAL group: F. Jansen, L. Duijnh, H. de Bruijn, I. Andersson, C. Behmer, K. Taylor, F. Kilburn-Toppin, M. van Goethem, R. Prevos, N. Salem, S. Paul. The authors thank the Medical Physics Department, Royal Surrey NHS Foundation Trust for the use of mammograms from the OPTIMAM Mammography Image Database funded by Cancer Research UK (C30682/A28396), and Sander van Woudenberg for all the help with the image processing.

Appendix A. Supplementary data

Supplementary material related to this article can be found in the online version, at doi:https://doi.org/10.1016/j.ejrad.2020.109464.

References

[1] S.H. Taplin, C.M. Rutter, C. Finder, M.T. Mandelson, F. Houn, E. White, Screening mammography: clinical image quality and the risk of interval breast cancer, Am. J. Roentgenol. 178 (2002) 797–803.
[2] S.A. Feig, Image quality of screening mammography: Effect on clinical outcome, Am. J. Roentgenol. 178 (2002) 805–807.
[3] G.W. Eklund, G. Cardenosa, W. Parsons, Assessing adequacy of mammographic image quality, Radiology 190 (1994) 297–307.
[4] C. Van Ongeval, H. Bosmans, A. Van Steen, Current challenges of full field digital mammography, Radiat. Prot. Dosimetry 117 (2005) 148–153.
[5] F. Zanca, J. Jacobs, C. Van Ongeval, F. Clauw, V. Gelis, C. Geniets, V. Provost, H. Pauwels, G. Marchal, H. Bosmans, Evaluation of clinical image processing algorithms used in digital mammography, Med. Phys. 36 (2009) 765–775.
[6] L.M. Warren, R.M. Given-Wilson, M.G. Wallis, J. Cooke, M.D. Halling-Brown, A. Mackenzie, D.P. Chakraborty, H. Bosmans, D.R. Dance, K.C. Young, The effect of image processing on the detection of cancers in digital mammography, Am. J. Roentgenol. 203 (2014) 387–393.
[7] R. van Engen, H. Bosmans, P. Heid, B. Lazzari, S. Schoop, M. Thijsen, K. Young, Digital mammography update. European protocol for the quality control of the physical and technical aspects of mammography screening, 51, part 2, European type testing, in: N. Perry, M. Broeders, C. de Wolff, S. Tornberg, R. Holland, L. von Karas (Eds.), Eur. Guidel. Qual. Assur. Breast Cancer Screen. Diagnosis, fourth ed. suppl., European Commission, Office for Official Publications of the European Union, Luxembourg, 2013, pp. 55–71.
