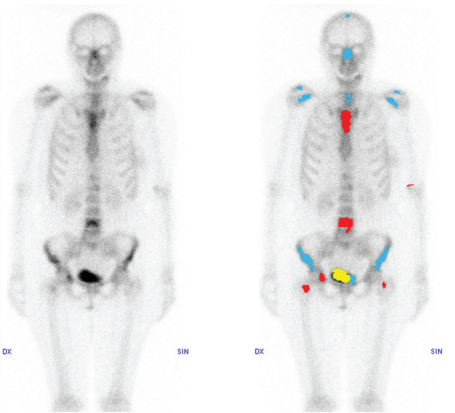
The Sahlgrenska Academy

Computer-Assisted Diagnosis for the Interpretation of Bone Scintigraphy

A new approach to improve diagnostic accuracy



May Sadik

Institute of Medicine
at Sahlgrenska Academy
University of Gothenburg

UNIVERSITY OF GOTHENBURG

Computer-Assisted Diagnosis for the Interpretation of Bone Scintigraphy

A new approach to improve diagnostic accuracy

May Sadik

UNIVERSITY OF GOTHENBURG

Department of Molecular and Clinical Medicine
Institute of Medicine
Sahlgrenska Academy at the University of Gothenburg
Gothenburg, Sweden

2009

The Swedish Medical Research Council and the Sahlgrenska Academy (ALF-grants) funded these studies.

Printed by Geson Hylte Tryck
Gothenburg, Sweden 2009

ISBN 978-91-628-7772-9 http://hdl.handle.net/2077/19652

To the memory of my grandmother Hagar

I will always remember the evenings we spent on your roof terrace,
breathing in the cool breeze of Baghdad,
listening to your radio,
lying close to each other and
watching the stars.

Ja visst gör det ont

Ja visst gör det ont när knoppar brister.
Varför skulle annars våren tveka?
Varför skulle all vår heta längtan
bindas i det frusna bitterbleka?
Höljet var ju knoppen hela vintern.
Vad är det för nytt, som tär och spränger?
Ja visst gör det ont när knoppar brister,
ont för det som växer
och det som stänger.

Ja nog är det svårt när droppar faller.
Skälvande av ängslan tungt de hänger,
klamrar sig vid kvisten, sväller, glider -
tyngden drar dem neråt, hur de klänger.
Svårt att vara oviss, rädd och delad,
svårt att känna djupet dra och kalla,
ändå sitta kvar och bara darra -
svårt att vilja stanna
och vilja falla.

Då, när det är värst och inget hjälper,
Brister som i jubel trädets knoppar.
Då, när ingen rädsla längre håller,
faller i ett glitter kvistens droppar
glömmer att de skrämdes av det nya
glömmer att de ängslades för färden -
känner en sekund sin största trygghet,
vilar i den tillit
som skapar världen.

Karin Boye 1900-1941

List of contents

Abstract

The most common cancers in the western countries are breast cancer in women and prostate cancer in men, and these cancer types, together with lung cancer, often metastasise to the skeleton. Bone scan is used to determine whether metastases are present, and the result of the examination serves as a guide in the choice of treatment strategy. Correct interpretation is, therefore, of great importance. The primary aim of this thesis is to investigate whether diagnostic accuracy of planar whole-body bone scan interpretations can be improved with the aid of a computer-assisted diagnosis (CAD_x) system. This is accomplished by four separate studies, of which the first shows that an automated CAD_x system is possible to develop for the interpretations of bone scans regarding the presence or absence of metastases. In the second study we investigated, in a nation-wide survey, the physicians' performance and the interpretive variations between readers for bone scan examinations. The physicians were asked to classify 59 images regarding the presence or absence of bone metastases. The images were selected to reflect the spectrum of pathology found in everyday clinical work. The physicians' interpretations were compared with final clinical assessment based on a 4.8 year follow-up period, and they were also compared pairwise with each other. The results showed a sensitivity ranging from 52% to 100%, with an average of 77% and a mean specificity of 96%. In addition, moderate agreement was found between readers. The experience from these studies resulted in the development of a second CAD_x system (third study) based on improved image processing and artificial neural network techniques and a larger database of whole-body bone scans. The CAD_x performance when tested on the 59 bone scans showed a sensitivity and a specificity of 90% and 89%, respectively. In the final study 35 physicians with varying levels of experience, working at 18 of the 30 nuclear medicine departments in Sweden interpreted the 59 bone scan images again, this time with the aid of the CAD_x system. The results showed a significant increase in sensitivity (88%) ($p<0.001$) without significant loss of specificity (94%). The area under the ROC curve increased from 0.925 without CAD_x to 0.961 ($p=0.005$) with CAD_x. The variation in interpretations decreased with CAD_x. In conclusion, a CAD_x system can improve diagnostic accuracy and reduce interpretive variations between physicians for bone scan examinations.

Key words: Diagnostic accuracy – Radionuclide imaging – Bone metastases – Breast cancer – Prostate cancer

Populärvetenskaplig sammanfattning

Bland de vanligaste cancerformerna i västvärlden är bröstcancer hos kvinnor och prostatacancer hos män och dessa cancerformer, inklusive lungcancer, sprider sig ofta till skelettet. Skelettscintigrafiska underökningar genomförs för att upptäcka tumörspridning och det är av stor vikt att bilderna tolkas korrekt för att patienten ska få rätt cancerbehandling. En korrekt diagnos är en av förutsättningarna för att framgångsrikt bota eller bromsa en aggressiv cancersjukdom. Dessutom är det viktigt att en undersökning bedöms på samma sätt oberoende av vid vilket sjukhus diagnostiken sker. Det primära syftet med avhandlingsarbetet var att undersöka om den diagnostiska säkerheten för tolkning av skelettscintigrafiska bilder kan förbättras med assistans av ett datorbeslutstöd. Detta uppnåddes med hjälp av fyra separata studier varav den första visade att det är möjligt att utveckla ett helautomatiserat datorbeslutstöd för tolkning av skelettumörfrågeställning. I arbete två undersöktes, i en nationell multicenterstudie, hur bra bedömare är på att hitta skelettumörer och hur eniga bedömarna är i tolkningarna när de jämförs parvis med varandra. Deltagarna ombads tolka 59 skelettscintigrafiska bilder avseende tumörer. Bilderna valdes så att de skulle representera den typiska fördelningen av tumörförekomst som påträffas i den kliniska vardagen. Resultaten visade att bedömarna hittade i genomsnitt 77% av patienterna med tumörer och 96% av patienterna utan tumörer. Bedömarna var dessutom måttligt eniga i tolkningarna när de jämfördes med varandra. Utifrån erfarenheten av dessa studier utvecklades ytterligare ett datorbeslutstöd (arbete III) baserat på förbättrad bildbehandling, förstärkt artificiella neurala nätverksteknik samt en större databas av skelettscintigrafiska bilder. När datorbeslutstödet testades på de 59 bilderna hittades 90% av patienterna med skelettumörer samt 89% utan tumörer. I det fjärde arbetet deltog 35 läkare från 18 olika sjukhus med varierande erfarenhet av skelettscintigrafisk tolkning. Läkarna ombads att återigen tolka samma 59 bilder, denna gång med assistans av ett beslutstöd. Läkarnas bedömningar jämfördes dels med den samlade kliniska slutbedömningen som bygger i snitt på 4,8 års uppföljningsperiod, dels jämfördes läkarnas tolkningar med varandra. Resultaten visade både att läkarna hittade fler patienter med skelettumörer (88%) och fick en mera samstämmig tolkning då beslutstödet användes. Sammanfattningsvis så kan ett datorbaserat beslutstöd assistera läkare till ökad diagnostisk säkerhet vid tolkning av skelettscintigrafiska bilder samt minska variationerna i tolkningarna mellan bedömare.

List of papers

This thesis is based on the following papers, which will be referred to in the text by their roman numerals:

I. Sadik M, Jakobsson D, Olofsson F, Ohlsson M, Suurkula M, Edenbrandt L. **A new computer-based decision-support system for the interpretation of bone scans.** *Nucl Med Commun.* 2006; 27:417-423.

II. Sadik M, Suurkula M, Höglund P, Järund A, Edenbrandt L. **Quality of planar whole-body bone scan interpretations-a nationwide survey.** *Eur J Nucl Med Mol Imaging.* 2008; 35:1464-1472.

III. Sadik M, Hamadeh I, Nordblom P, Suurkula M, Höglund P, Ohlsson M, Edenbrandt L. **Computer-assisted interpretation of planar whole-body bone scans.** *J Nucl Med.* 2008; 49:1958-1965.

IV. Sadik M, Suurkula M, Höglund P, Järund A, Edenbrandt L. **Improved classifications of planar whole-body bone scans using a computer-assisted diagnosis system: a multicenter, multiple-reader, multiple-case study.** *J Nucl Med.* 2009; 50:368-375.

Abbreviations

CAD	Computer-Assisted Detection or Computer-Assisted Diagnosis
CAD_e	Computer-Assisted Detection
CAD_x	Computer-Assisted Diagnosis
CT	Computed Tomography
DBMMRMC	Dorfman-Berbaum-Metz Multiple-Reader Multiple-Cases
FN	False-Negative
FP	False-Positive
κ	Kappa
MRI	Magnetic Resonance Imaging
PA	Percentage Agreement
RC	Relative Concentration
ROC	Receiver Operating Characteristic curve
RP	Relative Position
RV	Relative rank Variance

Background

Image interpretations

Rapid technological innovations in the imaging field, such as in functional and metabolic imaging, ultrasound, radiology and interventional radiology, contribute to seemingly unlimited horizons of diagnostic possibilities. However, with wider possibilities the complexity of the interpretation process rises and becomes an even more demanding task. Furthermore, the expanding and rapidly aging population requires an increasing volume of diagnostic examinations. Physicians and radiologists are, therefore, facing an increasing workload, but must still manage to read the diagnostic images carefully, avoiding errors in interpretation which may otherwise lead to an adverse effect in patient management. Surveys between 1999 and 2003 report an increase in radiologists' workload by 39%, measured in relative value units (RVU) (1). The RVU measures the physicians' productivity, taking into account the complexity of the examinations being interpreted and the percentage of time the average reader spends on the interpretations (2). As the volume of investigations increases, the risk of errors and the associated anxiety accelerates even more rapidly (2).

Most interpretations of diagnostic images are made visually, which makes the classifications observer-dependent. Factors that result in differences in interpretation are either personality dependent and/or random. One type of personality difference is when a reader has the tendency to report abnormal findings more often than another reader, who conversely has the tendency to reject findings, that is, the physicians systematically either overestimate or underestimate the significance of the findings. The other type of personality-dependent difference is when a reader is uncertain and often use words like "bone metastases cannot be ruled out with certainty" or "bone metastases probable", while another reader is more definite and often use statements like "absence of bone metastases" or "definite presence of bone metastases". Random errors, on the other hand, arise due to disturbing factors related to a busy practice, like loos of concentration or fatigue (3, 4). These types of image interpretation errors can in the worst cases lead to lack of appropriate treatment or the opposite, unnecessary additional examinations or treatment.

Fortunately, the majority of image interpretations are correct, but what can be done about those that are wrong or misleading? One way to assist the image reader could be by using a

computer system which alerts the reader to suspicious findings and propose a "second opinion" for the interpretation.

Decision support systems

The idea of a computer-assisted detection/diagnosis (CAD) system is to draw the physician's attention to pathological changes, while minimizing the risk of abnormal findings being overlooked (5-8). The general concept is to assist the reader by combining his or her competence and knowledge with the computer's capability to detect lesions in medical images. Computer alerts should not be considered surrogates or possible replacements for human experts; rather they should facilitate and complement their work.

CAD systems could be used either for the detection (CAD_e) of lesions by flagging for suspicious uptake in the images, or further developed also to deliver a diagnosis (CAD_x) for the whole examination. In either case, when CAD is used as a "second opinion", the physician makes the first decision by judging the images without CAD, then asks for CAD advice and ultimately makes the final interpretation. In some cases in which physicians are confident of their judgment, they may agree with the CAD output, or disagree and then disregard the CAD advice. In other cases, where the physician is less confident, the final decision can be improved by the use of a CAD advice, if of course the CAD advice is correct. The higher the performance of the CAD system, the better is the effect on the final interpretation. This approach does not, however, require the CAD system to be equal to or better than the physician but to complement the reader. CAD systems used as a "second opinion" place the responsibility for lesion detection on the physician, and they do not entirely rely on CAD. Used in this order, CAD constrains the physician to maintain a high level of knowledge and vigilance and then adds the alerting effect of the computer system.

Other ways of using CAD are either as a "first reader" or as a "concurrent reader". CAD as a "first reader" presents the diagnosis directly and the physician makes the final image interpretation by either accepting or rejecting the advice of the CAD. Used as a "concurrent reader", the images are displayed simultaneously on the computer screen. These last two approaches can have the advantage of shortening the interpretation time for each examination, but require the CAD system to achieve very high sensitivity in order not to miss a suspicious finding.

CAD systems can improve medical practice in many different fields, as has been shown in a recent systematic review of randomised controlled trials (8). It has been applied to all imaging modalities, including radiography, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound and nuclear imaging, used for all body parts such as the skull, thorax, heart, abdomen and extremities, and all kinds of examinations, including skeletal imaging, soft tissue imaging, functional imaging and angiography (5, 9). Currently, most CAD systems developed in the imaging field are for the detection of breast lesions in mammography screening (6, 10-13), the detection of lung nodules in chest CT (7), the detection of polyps in CT colonography (14-16), the detection of cervical cancers in cytology (17, 18) and for the diagnosis of ischemia and infarction in myocardial perfusion scintigrams (19, 20). Yet, previous work may be considered as an inception, since computer alerts may contribute towards important benefits in medical images in the future. As the physicians' interpretation time is limited, these kinds of systems can help to deal quickly with the constantly increasing number of images and the high flow of information.

Bone scintigraphy

The reason bone scintigraphy was chosen as the research field was because bone scans are the most frequently performed nuclear medicine examination, and are widely accepted as a method of choice for the initial diagnosis of bone and joint changes in patients with oncological diseases (21-25). The interpretation is, however, a difficult pattern-recognition task. Non-neoplastic diseases can also reveal abnormalities in the images, and a number of differential diagnosis and error sources should be considered (26). Some previous studies have shown a lack of sensitivity in the reporting of bone metastases, and a variation in interpretations that was substantial, even for such a well-established and widely used diagnostic method (27, 28).

Quantification programs presenting the extent of metastases in whole-body bone scans have been presented previously (29, 30). Their results showed a high correlation with corresponding visual or manually drawn region-of-interest analysis. These types of programs may be of value in clinical trials in order to present an objective measure of the extent of bone metastases, but they are probably too time-consuming for clinical use (30). The measuring time for the method presented by Noguchi and co-workers was on average 16

minutes per examination (30). The method by Erdi and colleagues is semi-automated and requires the user to insert a seed point into each metastatic region on the image before the system can demarcate a region of interest (29). Yin and Chiu developed a CAD_e system for bone scans in order to provide physicians with warning marks for suspicious radiotracer uptake, again without a presentation of the likely diagnosis for the whole examination (31). Their system showed a high detection rate, but the mean number of false-positive (FP) detections was 37 in an abnormal image and 46 in a normal image. Their system performed better for hands and legs and worse for the head and vertebrae regions, where metastatic bone disease is mostly located. CAD_e systems with lower FP detections or CAD_x with higher specificity, with completely automated analysis, and methods which propose recommendations for the possible diagnosis will most likely be required before physicians can adopt this type of technique as part of their clinical workflow.

Aims

The primary aim of this thesis was to investigate whether diagnostic accuracy of whole-body bone scan interpretations can be improved with the aid of a computer-assisted diagnosis system. This was accomplished by four separate studies with the specific aims to:

I. develop a completely automated method, based on image processing techniques and artificial neural networks, for the interpretation of bone scans regarding the presence or absence of metastases.

II. investigate, in a nationwide survey, the inter-observer variation and performance in interpretations regarding the presence or absence of bone metastatic disease based solely on bone scan images.

III. develop a completely automated computer-assisted diagnosis system for interpretation of bone scans regarding the presence or absence of metastases, based on improved image-processing, artificial neural network techniques and a large database of whole-body bone scans.

IV. investigate, in a multi-centre study, whether physicians benefit from the advice of our computer-assisted diagnosis system by reducing inter-observer variation and improving performance in the interpretations of bone scans regarding the presence or absence of bone metastases.

Development of a CAD_x system

An automated CAD_x system for the interpretation of bone scans regarding the presence or absence of metastases was developed. In order to create such a system a multidisciplinary research group consisting of physicians, engineers, physicists, a statistician and a technologist was established. The four important cornerstones in this thesis were image processing, machine learning, databases and clinical evaluation.

The CAD_x system used for the interpretation of bone scans consists of image processing techniques and a machine learning method called artificial neural networks (ANN). The image processing presents features from the bone scan images to the ANN, which are then used to interpret the examination. Since these methods are data-driven, i.e. learning by training, a database consisting of a large number of bone scans was used. After the development of the CAD_x system, the performance had to be evaluated by utilizing a separate test group, representing cases that are normally seen in daily clinical work.

The decision to adopt a new technology depends on adequate evaluation of the performance and usefulness of this method. The important issue is yet not the performance of CAD_x per se, but whether physicians benefit from it. As the CAD_x system is intended to be used as a "second opinion", the performance and usefulness of CAD_x is equal to the performance achieved by the physicians, who make the final decision, by using the CAD_x advice. The following study design, both for the development and testing of the decision support and for investigating the physicians' performance without and with CAD_x, is used in an attempt to evaluate the clinical value of CAD_x in the final decision-making.

Image processing

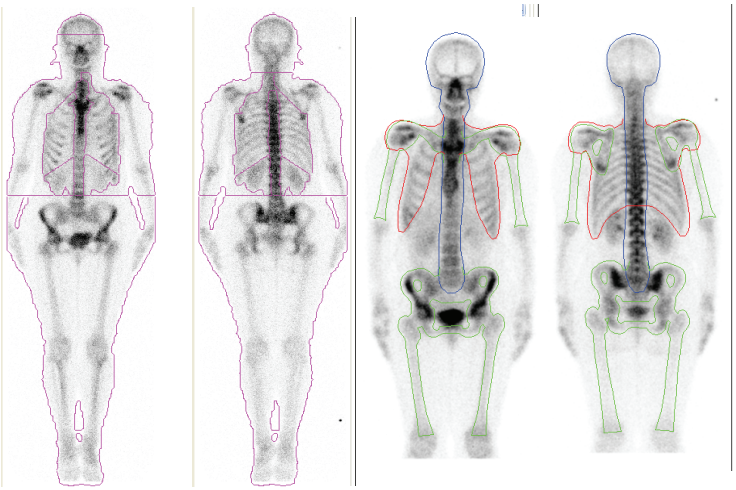
Each bone scan image, anterior and posterior, is composed of large number of picture elements (pixels) (256x1024) and each pixel is presented by a value, or "a variable". The total number of variables for the whole examination therefore exceeds 0.5 million ((256x1024) x 2). Only some of these variables represent the actual skeleton, while the rest are outside the body. The purpose of the image processing operation is to reduce the large number of variables by extracting relevant and excluding irrelevant information before processing it further to the ANN. Relevant information could, for example, be the number of hot spots present in the images or the distribution of the hot spots. Information excluded

could, for example, be symmetrical hot spots, a hot spot representing the bladder and the area outside the body. Image processing is one of the most important aspects of the development of a CAD_x system. Prior knowledge is needed of the examination that is to be interpreted, and the goal is to transform this knowledge into mathematical formulae and incorporate it into the image processing. There are no general rules as to how to extract relevant features, and this has to be tested for each individual problem.

The three basic operations used in the image processing are segmentation, hot spot detection and feature extraction.

Segmentation

The segmentation of the skeleton defines the region of interest. A precise segmentation makes it possible to use different variations of the algorithm for the detection of hot spots in different parts of the image, and to present information in greater detail regarding the localization and distribution of hot spots to the ANN. Two different approaches where used to segment the skeleton. In paper I thresholding was used to separate the bones from the background, which defined the outer contour of the body (Fig. 1a). In paper III active-shape models (ASM) were used (32), which is a statistical approach to find different skeleton parts (Fig. 1b). ASM allows expected variations in size and shape of the skeleton and considers grey-level appearance to find the borders of the skeleton. A robust segmentation is a prerequisite for successful hot spot detection, since the localization of each potential hot spot was obtained on the basis of the result of the segmentation process.



1a. 1b.

Figure 1. *The two different segmentation approaches used in paper I (1a) and paper III (1b).*

Hot spot detection

A region-specific threshold algorithm was used to detect potential hot spots. The algorithm was based on the mean and standard deviation of all pixel count values from a specific region. Clusters of pixels with count values above this threshold, with a cluster size exceeding six pixels in paper I and 13 pixels in paper III, were regarded as potential hot spots. A low number of pixels increases the risk of detecting noise, i.e. false-positive detections. By increasing the number of pixels in paper III, the false-positive detections decreased and consequently the false-positive interpretations. The localization of each potential hot spot could be obtained, based on the result of the segmentation process. Hot spots corresponding to the bladder and the kidneys were excluded based on location and size.

Feature extraction

Mathematical algorithms were developed to extract useful parameters from the segmented regions. The features were selected to describe both the hot spot itself and its relation to other hot spots and the surrounding region. The size, shape, intensity and localization of a hot spot were calculated, as well as the intensity characteristics of the region in which the hot spot was located. An important difference in image processing between the first and the second CAD_x was the feature extraction operation. Fourteen features were used in the first CAD_x (Paper I) to describe the anterior and posterior images, while this number was extended to 45

features to describe each hot spot in paper III, and additionally 26 features to describe the whole bone scan examination. The increased numbers of features enabled the image processing operation to better describe the characteristics of each hot spot, i.e. the size, shape, intensity and localization before presenting them to ANN. The resulting image features are used as input to ANN, used to interpret the hot spots and the complete examination.

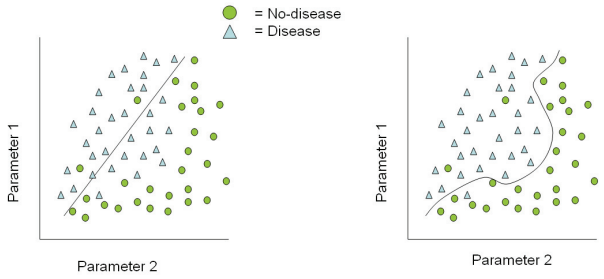
Machine learning method

Medical decisions are rarely based on a single measurement; instead physicians often consider several factors and parameters before making their final diagnosis. Physicians have learned to weight the importance of these factors or measurements through acquired knowledge, experience from previous cases, discussions with senior experts and comparison with the final clinical outcome. In order to shorten this long learning curve and minimise variations in interpretations due to turnover of staff, decision support systems have been developed to assist physicians by providing knowledge based on a large number of cases.

There are different methods used in the development of a decision aid. Expert systems based on rules predefined by physicians (33), standard statistical methods (33) or artificial neural networks (34) are some methods used. Which method to use depends on the problem that has to be solved.

Interpretation of bone scan images is a complex process based on pattern-recognition weighting different variables, like homogeneity of the radiotracer uptake, symmetry, localization, distribution, intensity, coverage, shape of hot spots etc. Furthermore, the separation between the two groups “disease” or “no disease” when classifying cases from a diagnostic test is often non-linear. Therefore, using an interpretation method assuming a linear boundary in an attempt to discriminate between these two populations would be less powerful (Fig. 2a). The advantage of the ANN is that they can be trained and adjusted to discriminate between “disease” or “no disease” with non-linear decision boundaries. (An illustration of a non-linear decision boundary can be seen in figure 2b). A prerequisite is a large, representative training group with the desired interpretation (gold standard). ANN are not programmed and are not, therefore, restricted to a set of predefined critical variables for their interpretations: rather they learn from examples, in the way people do. The ANN can

generalize by detecting similarities between new patterns and previously stored patterns and are, therefore, ideal for complex pattern recognition. Another set of cases, a test group, with a gold standard will be needed to test the performance of the networks.



2a. 2b.

Figure 2. *Using a linear boundary in this population for the separation between "disease" or "no disease" (2a) is not as powerful as the non-linear discrimination boundary (2b).*

The probably most used ANN design for interpretation problems, such as discriminating between "disease" or "no disease" is the multilayer perceptron (35). The most common construction has three different layers; one input layer, one hidden layer and one output layer (Fig.3). Each layer consists of one or more processing units called nodes (mimicking neurons in the human brain). The nodes are interconnected by a set of "weights" (analogous to synaptic connections in the nervous system) allowing signals to travel through the network (Fig.3).

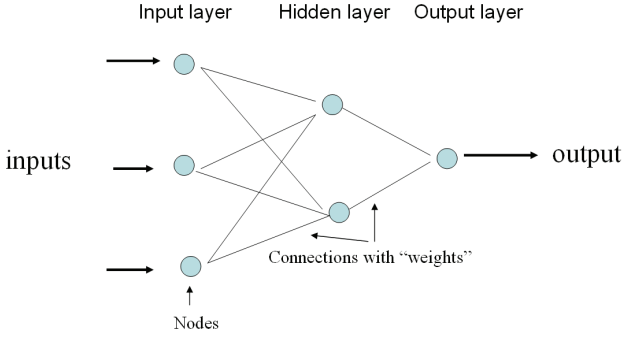


Figure 3. *Example of a simple three-layer ANN. The resulting image features from the image processing operation are presented as inputs, processed through the network, and a likelihood value for a patient of having bone metastases is finally delivered from the output layer.*

The input layer contains the same number of nodes as the number of input variables (features). (For example, there are 14 nodes in the input layer (Paper I), one for each feature used). The output node computes the likelihood of the event, e.g. the likelihood of having bone metastases. The number of nodes used in the output layer depends on how many categories the interpretation problem requires. Every node in a layer is connected with each node in adjacent layers. Each node sums the weighted signals it receives from its input connections and produces an output signal which is a non-linear function of this input. This mimics the way that incoming nerve impulses are aggregated in a biological neuron which will fire (an "action potential") if these signals exceed the activation threshold. The computational power in an ANN lies in the interconnections between the nodes containing the weights, together with non-linear activation functions. Weights are adjusted in the training process until the output from the network better agrees with gold standard. This is the "memory" of the ANN. Once the training stops the memory cannot be modified until a new set of training cases are presented and weights adjusted.

In the first CAD_x system (Paper I) one set of ANN was used as an interpreter for the whole examination, while in the second CAD_x (Paper III) two sets of ANN were used; the first interpreting each hot spot and the second interpreting the whole bone scan examination. In paper III the second network weighted the most important hot spot information received from the "hot spot" network, before interpreting the whole bone scan images. By this approach, increased radiotracer uptake caused by, for example, bad teeth or sternotomy was flagged as suspicious findings but was not interpreted as bone metastases in the final classification made by CAD_x (Fig.4). Using two sets of ANN enabled the program to increase the specificity in the final interpretation for the whole examination.

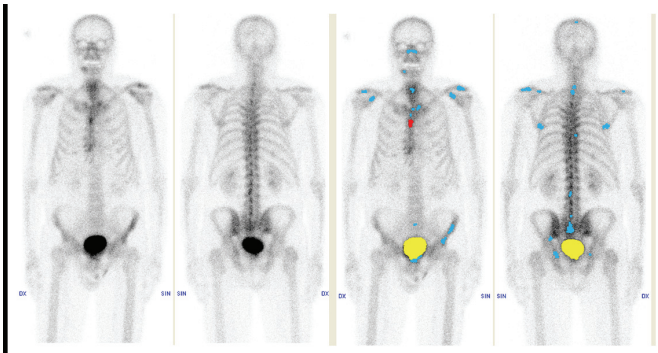


Figure 4. *A 70-year-old man with prostate cancer. Increased radiotracer uptake can be seen in the right part of the mandible, most probably due to bad teeth, and in the sternum secondary to sternotomy. The artificial neural networks classifying hot spots separately indicate that uptake in the sternum could be a metastasis, but the neural networks classifying the complete examination, considering all hot spots, correctly report 'absence of bone metastases'. Suspicious metastases are marked in red, symmetrical or benign radiotracer uptake in blue and the bladder is yellow.*

Databases

Images

Ideally the databases used for CAD development and evaluation should be at least of the order of hundreds, including representative cases found in a clinical routine. Researchers who have developed automated methods for the interpretation of coronary artery disease in myocardial Bull's-Eye scintigrams have included the same 135 patients both for training and evaluation, by using the "leave-one-out" validation procedure (19). One patient examination was used as an evaluation case, whereas the remaining 134 patients were used for training. This procedure was repeated 135 times, such that each case in the data set was used as a test case once (19). Other CAD_e developers for programs used for the detection of polyps on CT colonography have used 141 cases as a training group and 119 in test (14), or 239 cases for training and "leave-one-out" for the validation (16), respectively, collected from different centres.

Yin and Chiu developed a CAD_e method providing physicians with warning marks for suspicious radiotracer uptake on bone scans, and used 20 selected patients in the training phase and further 109 patients from the same database for the test group (82 abnormal and 27 normal images) (31). In our first decision-support system (Paper I), 200 consecutive patients,

who had undergone whole-body bone scintigraphy due to suspected bone metastatic disease were included. We chose a consecutive approach in order to train and test the CAD_x with representative patient material which is normally seen in the clinical routine, e.g. prevalence of bone metastases, age, gender etc. The total group was then randomly divided into 100 cases each for the training and testing of the CAD_x system (Table 1). The number of cases used in paper I lie in the range of what has been used by others who have developed successful decision support systems. Our approach was also superior to those used by Lindahl et al and Halligan et al, as we included two different patient groups for training and testing (16, 19).

A major step forward in the improvement process between the first (Paper I) and second CAD_x (Paper III) system is that the number of training cases was increased from 100 to 810 consecutive patients (Table 1). This was done in order to present an even larger variation of bone scan findings, and thus enable the second CAD_x system to better differentiate normal or benign uptake patterns from pathological processes. The reason why we included this large number of cases was because it is the recommended number of bone scans a physician should interpret during his or her specialist training.

For the testing of CAD_x (Paper III) a group of 59 patients were retrospectively selected, who had undergone whole-body bone scintigraphy due to suspected bone metastatic disease, and who had had at least one follow-up bone scan (Table 1). These patients had a similar distribution of age, gender and prevalence of bone metastases to the consecutive material (Table 1). The reason why the number of test cases was reduced between the first and second CAD_x is because these 59 patients were not only used as the validation group for CAD_x but also in the evaluation of the physician's performance and inter-observer variation (Paper II and IV). A high number of patient studies makes it difficult to recruit physicians, due to the amount of time it would demand from each. We therefore decided to include as many cases as it would be possible to read in approximately 1 hour.

Table 1. Study population in paper I and III.

	Training group (paper I)	Test group (paper I)	Training group (paper III)	Test group (paper III)
No. of patients	100	100	810	59
Female (%)	31	32	35	31
Mean age, years (range)	68 (40–88)	69 (36–89)	66 (25–92)	65 (43–86)
Prevalence of bone metastases (%)	36	31	34	36

Gold standard

Different alternatives can be used as gold standard for the diagnosis of malignant disease, of which histological verification is considered as the most accurate. Biopsy is possible to use if included as part of the routine workflow as, for example, when breast cancer is suspected on mammography, or for the analysis of malignancy in polyps detected on colonoscopy. This type of gold standard was used in the evaluation of CAD_e programs in the fields of mammography and CT colography (10-13, 15, 16). However, biopsy was only performed when malignancies were suspected, while in the remaining cases interpreted as normal or benign, an expert physician or panel of physicians who read the images served as the reference standard. One should also bear in mind that biopsy has its limitations. Layfield reported in a recent review that the diagnostic accuracy in fine-needle aspiration for bone lesions is 85%, and that the insufficiency rates could vary from 4% to 33% (36).

Using biopsy as a gold standard is impossible in practice for each hot spot found in the bone scans, as multiple skeletal parts could be involved. In such circumstances acceptable and frequently used reference standards are: a human expert, consensus agreement between a panel of experts, application of a majority rule from an expert panel, or, taken even further, using final clinical outcome based on a long follow-up period.

The gold standard interpretation of the patients in the training and test groups (Paper I) regarding presence or absence of bone metastases was based on clinical reports and bone scan images. These reports and the corresponding bone scans were re-evaluated by an

experienced physician and a trained technologist, who estimated the probability of bone metastases on an analogue scale from 0 to 1.

In paper III during the improvement of the second CAD_x system, the gold standard for the training group was further strengthened. In difficult cases results from other diagnostic examinations of the patient, for example, follow-up scans, MRI, X-rays or CT when available, were considered in the re-evaluation.

In order to come as close as possible to the absolute truth regarding the presence or absence of bone metastases in the 59 patients, who were used both as the test group for CAD_x (Paper III) and in the evaluation of the physician's performance and inter-observer variation without and with CAD_x (Paper II and IV), we used final clinical assessment made by an experienced physician. This assessment was based on all the bone scan images, including the follow-up scans, the patient's computerised medical record including the results of laboratory tests, and all available diagnostic images (MRI, CT and X-ray) for a mean follow-up duration of 4.8 years. The aim of this approach was to use all the available patient data.

Using follow-up as a gold standard is preferable compared with consensus agreement among an expert panel, basing their interpretations only on the bone scan images. Just because experts agree in their interpretations, this does not automatically mean that their interpretations are correct. However, final clinical assessment was carried out by one expert physician, which could be seen as a limitation. An expert panel discussing all cases during the follow-up would have further strengthened the reference standard.

If we had used a panel of experts as the gold standard, and the majority rule had been applied to the 11 experienced physician interpretations (of the total 35 readers, paper IV), then 53 of the 59 cases would have had the same classification regarding presence or absence of bone metastases. (These participating physicians had access to the 59 current patient examinations but not to the follow-up information.) In the 6 cases that were diagnosed differently from our gold standard, follow-up had an important impact on the final diagnosis. In 2 of the 6 cases nearly all (10/11) the experienced physicians classified the increased uptake as false-negative. One of these patients had a positive biopsy, and the other showed increased uptake in lesion size and intensity on the follow-up scans. In the remaining 4 cases there were more disagreements among the 11 experienced physicians, 3 cases with 6 votes against 5, and one

with 7 votes against 4. This indicates that our gold standard based on final clinical assessment is less dependent on the experience of this one expert physician.

Clinical evaluation

Clinical evaluation should ideally be designed so that the results also have implications for physicians and patients other than those who participated in the trial. The study design used is referred to as "three factors" – multiple-reader, multiple-cases and multiple-modality, "fully crossed" design – all physicians interpreted all images twice, once without and once with CAD_x. The goal is to have as many readers and cases as possible. However, there is a trade-off between the number of physicians who will agree to participate in quality assessment studies and the number of cases each physician is asked to read. Authors who have used the same study design to investigate the physicians' performance without and with CAD usually included few (three to ten) readers, some of them working at the same hospital, and in some studies only experienced readers were selected to join (14-16, 19, 20). In these trials the physicians were asked to read between 30 to 135 cases. In contrast to the previously mentioned studies, we invited all the physicians in Sweden interpreting bone scans as part of their daily routine to participate in the nation-wide survey (Paper II). Thirty-seven physicians (of an estimated 100–125 physicians), with various levels of experience, from 18 of the 30 nuclear medicine departments in Sweden, agreed to participate. The physicians were instructed to visually review 59 whole-body, anterior and posterior, bone scan images without CAD_x and – one year later – with CAD_x. Thirty-five physicians participated on both occasions.

In the field of mammography, researchers have used different study designs when investigating the effect of CAD_e on physicians' performance. In two prospective studies 12,860 (10) and 8,682 (12) cases were included, respectively, and interpreted by one of two and one of seven physicians, i.e. not all physicians interpreted all cases. Their CAD_e systems were already implemented and part of the routine workflow. The need for inclusion of thousands of cases in these studies is due to the fact that the incidence of breast cancer in a screening population is low. Only 3 to 10 cancers are diagnosed out of 1,000 women screened (12). Therefore, in order to detect a significant positive effect using CAD_e a high volume of

cases is needed. In the current material, on the other hand, 36% of the 59 test cases had bone metastases (Table 1).

In a large Danish study, Rossing et al investigated the quality of reporting of bone scans from 842 breast cancer patients collected at 12 centres (27). Again, not all physicians interpreted each of the 842 scans, but each of the 12 centres contributed with some of these patients.

In order to be able to compare the physician's performance and inter-observer variation both without and with CAD_x the scale given bellow was used in the interpretation of the presence or absence of bone metastases. Commonly in their routine clinical work physicians use different statements in their reporting; some mention certain findings while others ignore them. These types of statements are, however, difficult to compile between the physicians and between two different occasions. Therefore, the physicians were asked to use a 4-point interpretation scale. A drawback could yet be that the physicians were constrained to a scale they might not be used to.

Grade 1: Absence of bone metastases

Grade 2: Bone metastases cannot be ruled out with certainty

Grade 3: Bone metastases probable

Grade 4: Definite presence of bone metastases

Results

The quality of the interpretation of bone scintigraphy, carried out by physicians or CAD_x regarding metastases, can be described using measurements of sensitivity and specificity (performance). One way of determining the physicians performance or the CAD_x performance is to classify the patients into two groups ("no bone metastases" or "bone metastases") according to the results of a gold standard and to compare the physicians' or the CAD_x interpretations with this reference standard. An alternative approach, which does not require a gold standard, is to measure the physicians' reproducibility by comparing them with each other (inter-observer agreement).

Performance

Paper I

The results from the first CAD_x system showed that it is possible to develop an automated method for the interpretation of whole-body bone scintigraphy regarding the presence or absence of bone metastases. When compared with gold standard (expert physician), the interpretations by the CAD_x system showed a sensitivity of 90% in the test group. CAD_x could correctly interpret 28 of the 31 patients with metastases. A false-positive interpretation of metastases was made in 18 of the 69 patients not classified as having metastases by the experienced physician, which resulted in a specificity of 74%. In order to determine whether these results are satisfactory and whether readers can be helped by such a system, our next aim was to identify the strength and limitations in the physicians' reporting, and to optimize a CAD_x system with the ability to complement the readers and prevent errors, with the final goal of increasing diagnostic accuracy.

Paper II

In the nation-wide survey 37 physicians interpreted 59 bone scan images and these were compared with final clinical assessment as gold standard, based on follow-up scans, the patient's computerized medical record including results of laboratory tests and all the available diagnostic images (MRI, CT, X-ray) for a mean follow-up period of almost 5 years. The physicians' results showed sensitivities ranging between 52% and 100%, with an average of 77%, indicating that the physicians either failed to detect the lesions or interpreted metastatic disease as benign findings. The specificities for the physicians were high; ranging between 79% and 100%, with an average of 96%. Rossing et al. investigated the quality of

reporting in bone scans by comparing the initial interpretations with a panel of three physicians as the gold standard (27). They showed a sensitivity and a specificity of 78% and 84% respectively. The difference in gold standard between the studies may partly explain the differences in specificities. Peters et al. studied clinical audit in nuclear medicine and showed that 19 reports out of 220 (8.6%) were classified as having non-trivial errors in the interpretations, that is errors with the potential adversely to influence patient management (28). Our results and those found by others indicate that the main problem in the interpretations of bone scan images seems to be false-negative errors. Among patients suffering from bone metastases the aim is to avoid skeletal-related events such as bone pain and pathological fractures, which is why this kind of misinterpretations should be reduced.

Paper III

The results from the first CAD_x system were encouraging, but further improvements were needed in order to apply the system in a day-to-day clinical setting. These experiences led to the development of a second method, based on improved image processing and artificial neural network techniques and a larger database of whole-body bone scans. An important improvement between the first and second CAD_x is that the specificity was increased from 74% to 89%. True negative interpretations were made for 34 of the 38 patients classified as not having bone metastases by the gold standard (follow-up). The second CAD_x system made correct interpretations for 19 of the 21 patients with bone metastases, showing the same level of sensitivity (90%) as in the first CAD_x.

It is difficult to make a direct comparison between our computer system and that of Yin and Chiu (31), as their system was used to provide warning marks to direct the physician's attention to suspicious radiotracer uptake, and not to provide a second opinion regarding the interpretation of the complete image, as in our system. They present per-lesion results while we present per-patient results (31). Their system showed a sensitivity in the per-lesion analysis of 91.5% and a mean false-positive detection of 37 per image in the whole material, i.e. a total of 4065 FP marks in 109 images. The problem of low specificity seems to be much less with our system. Our sensitivity and specificity were also higher than those presented by Sajn and co-authors (37), 79.6% and 85.4%, respectively, who presented an automatic method for analysis of whole-body bone scans. One explanation could be that we included a larger number of patients in the training process and used different techniques in the development of the system.

Ellis and co-workers compared two CAD_e systems (R2 ImageChecker M1000, version 5.0A and iCAD Second Look, version 6.0 mid operating point) used in mammography for the detection of breast cancer (38). The per-study sensitivity and per-study specificity for R2 ImageChecker and iCAD Second Look were 82% and 39%, and 61% and 31%, respectively. Both CAD_e systems are optimised towards higher sensitivity at the cost of much lower specificity. A low specificity implies that the physicians have to investigate many false-positive marks carefully, which can reduce the readers' acceptance of the system. Despite the fact that neither our system nor theirs is meant to be used independently, a high specificity is required in order not to increase interpretation time and induce the need for additional investigations.

Paper IV

The important issue is, however, not the performance of CAD_x per se, but whether physicians benefit from it. We found that, when the 59 images were interpreted with the aid of the CAD_x system, the results showed a positive additive effect on the 35 physician's performance. The combination of the physician's high specificity with the high sensitivity of CAD_x resulted in significantly increased sensitivity, from 78% without CAD_x to 88% ($p<0.001$) in detecting bone metastases, without significant loss of specificity. The following two cases illustrate the synergetic effect between CAD_x and the readers. The case shown in figure 5 was correctly classified as metastases by the CAD_x system, and with the computer's advice 30 physicians, instead of 17 without CAD_x advice, made a correct interpretation. Figure 6 shows an example of a patient with fractures, misinterpreted by the CAD_x system as having metastases, but correctly classified by 33 out of 35 physicians, despite false-positive CAD_x advice.

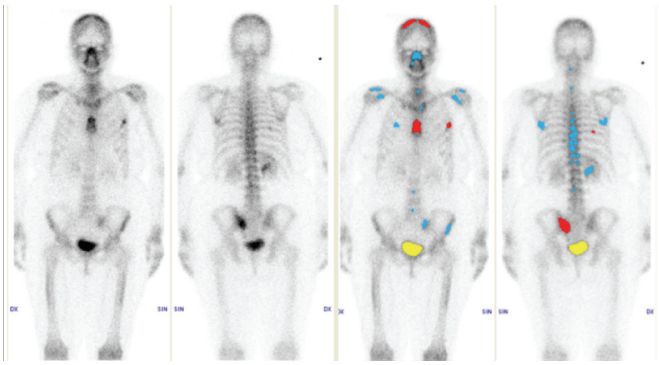


Figure 5. *With the advice of the CAD_x system 30 of the 35 physicians classified this case as a true-positive compared with 17 physicians without the CAD_x advice. A 47-year-old woman with breast cancer. Multiple focal increases in pathological radiotracer uptake can be visualized. X-ray of the left thorax verifies bone metastases, and the medical record stated metastases in the bone, liver and lungs. Suspicious metastases are marked in red, symmetrical or benign radiotracer uptake in blue and the bladder is yellow.*

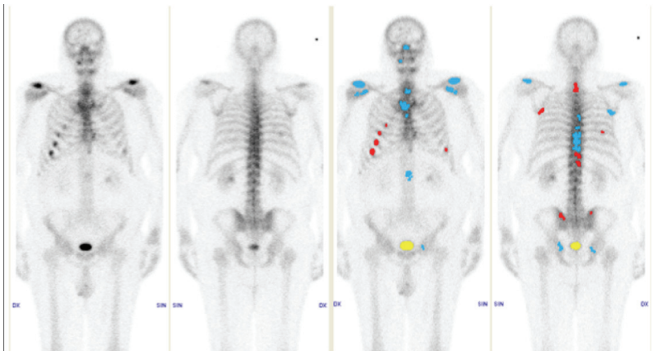


Figure 6. *A 63-year-old man with prostate cancer. Increased radiotracer uptake can mainly be seen in the rib/costa cartilages, which disappeared on the follow-up scan. The localizations speak in favour of fractures. The CAD_x system classifies this patient as having metastases, but the majority of the physicians (33/35) interpreted the images correctly, despite erroneous advice from the CAD_x system. Suspicious metastases are marked in red, symmetrical or benign radiotracer uptake in blue and the bladder is yellow.*

Several studies in different medical fields have been published highlighting the positive effect of CAD on physician's detection of abnormalities, and also showing less variation in the reporting. In CT colonography, CAD_e significantly increased the detection of polyps, by an average of 9.1%, especially in those of small and medium size (16). Taylor et al reported that,

despite false-positive CAD_e marks, the system did not adversely influence correct reader interpretation (15). Furthermore, Baker and colleges demonstrated that CAD_e can be useful for less-experienced readers, and showed that the average sensitivity to polyp detection significantly increased from 81% to 91% with CAD_e (14).

When interpreting diagnostic images, the final report should preferably, clarify whether or not the patient has bone metastases. However, in actual everyday practice these definite diagnosis are not always possible due to difficult, uncertain findings, and more than two categories are usually used to describe the spectrum of cases seen. The physicians were, therefore, asked to interpret the bone scan images regarding the presence or absence of metastases using a 4-categorical scale. The performance of each physician was measured as the area under the receiver operating characteristic (ROC) curve. The ROC area can present a value between 0 and 1, with 1 indicating perfect performance. The differences in ROC areas without and with the CAD_x system for all physicians were calculated using a multireader-multicase ROC analysis of variance trapezoidal area analysis (DBMMRMC 2.2) (39-46). This ROC analysis considers cases and readers as random samples. This means that the analysis is not only meant to reflect the performance in the particular sample studied, but provide an estimate of performance "on average" in similar cases and physicians and readings that were not studied. The results can, therefore, be generalized to both the population of cases and that of readers from which the test samples of cases and readers were drawn (39-46).

When the physicians considered the computer advice in their interpretations, the area under the ROC curves rose significantly from 0.925 without to 0.961 ($p=0.005$) with CAD_x. These findings are in accordance with what others have reported (19, 20) using computer advice as a second-opinion in the interpretations of myocardial perfusion scintigrams. Researchers have found improved performance, expressed as increases in the areas under the ROC curves from 0.65 to 0.70 and from 0.79 to 0.82 for two vascular territories when physicians used CAD_x (19). Others in the same field also showed benefits from CAD_x, expressed as increased sensitivity for ischaemia from 81% without to 86% with the CAD_x system (20).

In the United States single reading is standard practice in mammography screening, but double reading has been used in some centres in order to increase sensitivity in the detection of breast cancer. This procedure is time-consuming and costly from a radiologist manpower perspective. Gromet investigated the efficacy of single reading (without CAD_e) compared to

double reading and to single reading with CAD_e and found that the sensitivity increased from 81.4% to 88.0% to 90.4%, respectively. Gromet concluded that CAD_e enhances the performance of a single reader, yielding increased sensitivity, with only a small increase in recall rate (i.e., the rate at which mammographically screened women are recalled for additional assessment). They have routinely converted from double reading to single reading with CAD_e in screening mammograms (6). Similarly, in a large, prospective, multi-centre study Gilbert and colleges found no significant difference between double-reading and single-reading with CAD_e, with a modest increase in recall rate (47). Furthermore, in a recent review presenting CAD systems in medical imaging, Doi summarized the results of six prospective studies on the usefulness of CAD_e for the detection of breast cancers in screening mammography. All six studies indicated an increase in the detection rates from 1.7% to 19.5% in finding cancers (5). Importantly, Warren Burhenne et al showed that 67% (286 of 427) of breast cancers were visible retrospectively in prior mammograms, and by using a panel of radiologists 27% (115 of 427) could have been diagnosed earlier (11). Their CAD_e system could successfully detect 77% of these 115 cases and would have alerted the physician.

In contrast to the majority of studies dealing with CAD_e in mammography screening, Fenton et al concluded that CAD_e is associated with reduced accuracy of interpretation. Their study showed an increase in sensitivity from 80.4% to 84.0% and an increase in cancer detection rates from 4.15 to 4.20 per 1,000, but these two findings were not statistically significant. The authors argued that CAD_e provides no benefits, because the study showed a 19.7% increase in biopsy rate, which was statically significant (48).

In the detection of pulmonary nodules using CT, Golding and co-workers summarized in a review the current state of the art regarding CAD techniques, and concluded that CAD serving as a second reader, may provide better sensitivity for small nodules, easier enumeration and better documentation, improved inter-observer and inter-scan consistency in follow-up examinations, and a more objective assessment of significant temporal change in lesion size and number (7). These examples of previous work and our multi-centre study (Paper IV) demonstrate that CAD could increase diagnostic accuracy and reduce interpretive variations in diagnostic images.

Percentage agreement (PA) was used in paper II-IV in order to calculate exact agreement between categorical assessments made by physicians or CAD_x, compared with a gold standard, or between paired physicians. A value of 100% implies perfect agreement. Thus, some agreement between two physicians could be expected to be caused by guessing and in order to investigate the agreement beyond that expected by chance, the kappa (κ) coefficient was used. A maximum value of 1 means perfect agreement.

In comparison with the gold standard, when classifying the 59 bone scans without CAD_x the 37 physicians showed a mean PA of 66% and a mean κ value of 0.50, indicating moderate agreement. The second CAD_x system performed PA and κ values of 76% and 0.58, respectively, when interpreting the same scans. Finally, when the 35 physicians classified the images with the advice of CAD_x they showed significantly increased PA (73%) ($p=0.00004$). The same trend was observed with κ, that the mean value rose significantly to 0.58 ($p=0.0001$) with CAD_x advice.

Inter-observer variation

Paper II and IV

Pairwise agreement between the 37 readers was analyzed by creating all combinations of physician pairs. On average, PA between the paired readers when interpreting the images without CAD_x was 64% and mean κ was 0.48, indicating moderate agreement. One year later 35 of these physicians interpreted the images again, this time with CAD_x assistance. The influence of the CAD_x system on inter-observer variation was studied in the following way: for each pair of physicians, their agreement was quantified both without and with CAD_x. Mean changes between these two occasions were calculated for all pairs, but no significance tests were applied, because the physicians were dependent on each other.

The results showed that most pairs increased their agreement (PA and κ) with the advice of CAD_x. This indicates that, when the physicians considered the CAD_x advice in their classifications, the inter-observer variation decreased, that is most pairs agreed more in the interpretations.

Systematic or random variations papers II-IV

κ statistics is commonly used in these kinds of studies. However, there are several disadvantages with κ. First, one should not compare the value of κ from different studies where the prevalence of the categories differs. Second, the κ value depends on the number of categories used. The more categories, the higher is the chance of disagreement between physicians and vice versa. Third, κ calculations merely indicate that there can be some disagreement, without any further explanation of the type of disagreement. Therefore, to overcome these shortcomings we used the non-parametric approach of Svensson and Holm, which allows a deeper analysis of the nature of performance or inter-observer variation (3, 4). Disagreement between two physicians could be systematic and/or random. Two types of systematic variations are possible - the first due to overestimation or underestimation of the interpretations, and the second due to concentration of the interpretations. Systematic overestimation occurs when one reader classifies cases as being more abnormal than does another reader, or, conversely, when this is a case of systematic underestimation by the other reader. Systematic concentration occurs when one reader uses the middle section of the 4-point scale ("cannot be ruled out" or "probable") more often than another reader, who uses the grades "absence" or "definitely bone metastasis" more often. Overestimation or underestimation is reflected by the variable relative position (RP), and concentration by the variable relative concentration (RC). The possible values for RP and RC range from -1 to 1, and a value of 0 indicates that no systematic disagreement is present. The pattern of random errors was quantified using the variable of relative rank variance (RV). Random errors could be caused by guessing, or losing concentration. The possible values for RV are between 0 and 1, with 0 indicating no random contribution.

The main reason for disagreement between the physicians and the gold standard when interpreting the images without CAD_x was that the physicians concentrated more on the central sections of the 4-point scale and used words like "bone metastases cannot be ruled out with certainty" or "bone metastases probable" (Grades 2 and 3), in contrast to the gold standard.

However, the contribution of systematic variations in position (RP) and concentration (RC) was small for the second CAD_x system, that is, the computer neither overestimated or underestimated the interpretations, nor concentrated the interpretations to a certain section of the 4-point scale compared with the gold standard.

When the physicians interpreted the images with CAD_x, the disagreement in concentration was still present but significantly reduced ($p=0.00002$), that is the physicians used the more uncertain middle section of the 4-point scale ("cannot be ruled out" or "probable") less often. These results show that a CAD_x system can influence the interpretive style of physicians, but I acknowledge that there can be different opinions regarding how frequently uncertainty should be indicated in clinical reports. In addition, the contribution of random errors (RV) caused by, for example, losing concentration, reading fatigue or interruptions during the interpretations, decreased significantly with CAD_x ($p=0.01$).

Conclusions

The first study demonstrates that a completely automated method for the interpretation of bone scans can be developed regarding the presence or absence of metastases.

The second study shows that false-negative errors were the major problem in the interpretations of bone scan images, while the specificities for the physicians were high. Furthermore, moderate inter-observer agreement was found when physicians were compared pairwise with each other. The reason for the disagreement was mainly systematic, but random variations also contributed.

The third study demonstrates that an automated CAD_x system, based on improved image-processing, artificial neural network techniques and a large database of whole-body bone scans could further amend the performance of the computer method, expressed as increased specificity with maintained high sensitivity.

The last study shows that a CAD_x system can improve the physician's sensitivity in detecting metastases and reduce inter-observer variation in planar whole-body bone scanning. The CAD_x system appears to have significant potential in assisting physicians in their clinical routine.

The final conclusion of this thesis is that the diagnostic accuracy of planar whole-body bone scan interpretations can be improved with the aid of a computer-assisted diagnosis system.

Summary

What was already known on this topic is that few computer-based decision aids are used routinely in health care, despite several reports of potentially valuable new methodologies. The decision to take up a new technology depends on adequate testing to prove that it is a substantial improvement over unsupported human decisions, and there are only few studies of that type. The primary aim of this thesis was to investigate whether diagnostic accuracy of whole-body bone scan interpretations can be improved with the aid of a computer-assisted diagnosis system.

We have demonstrated that an automated CAD_x method can be developed regarding the presence or absence of bone metastases. In order to investigate the quality of reporting and estimate the need for such a decision support system, a nation-wide survey was carried out. All the physicians in Sweden interpreting bone scan images were invited to participate. They were asked to interpret bone scan images reflecting the spectrum of pathology found in every day clinical work. This is, to our knowledge, the most extensive investigation of the quality of reporting in this field. The results showed that the main problem in the interpretations of bone scan images was due to false-negative interpretations and moderate agreement was found between the readers when compared with each other. The experience from these studies resulted in the development of a second CAD_x system. By several efforts, such as improved image-processing and artificial neural network techniques and a large database of whole-body bone scans a robust, completely automated CAD_x system that propose recommendations for a possible diagnosis was developed and sent to the physicians who participated in the nation-wide survey.

The purpose of a decision support system is to aid physicians and not to replace them. The idea is to draw the physician's attention to abnormalities that might pass undetected. The physician is still responsible for lesion analysis and the final interpretation of the examinations. Our final study has demonstrated that this collaborative effect, obtained by combining the physician's competence with the computer's capability in detecting lesions, has resulted in increased sensitivity, maintained high specificity, and reduced inter-observer variation in planar whole-body bone scanning. These multiple-reader, multiple-case studies, comparing unaided to aided performance of multiple readers from multiple hospitals, may serve as an example of the proper evaluation of a decision aid. This thesis demonstrates the

development of an idea that ended up as a programme tested by 35 physicians working at 18 different hospitals in the country.

Other applications of the automated decision support system might be to highlight suspicious uptake for technologists during the acquisition of the images, so that they can consider obtaining additional views. It can also be used for educational purposes, to facilitate the understanding of bone scan interpretations for physicians, nurses, technologists, physicists and students and to shorten the learning curve needed to achieve high-quality reports.

Future developments in this field will focus on:

- quantification of the extent of bone metastases
- monitoring of progression/regression of the tumours on follow-up scans
- alerting for poor image quality due to low count rate

A phantom study is planned in order to determine whether the CAD_x system shows the same performance on images acquired with different gamma cameras or different protocols, for example when using faster scanning during the image acquisition or injecting lower radiotracer dose to the patient.

Ethics

The studies were approved by the Research Ethics Committee at Gothenburg University.

Conflict of interest

Lars Edenbrandt, Andreas Järund and Mattias Ohlsson are employed by and are shareholders in EXINI Diagnostics AB, which provides CAD_x software for nuclear medicine studies. Iman Hamadeh and Pierre Nordblom were employed by EXINI Diagnostics AB. The company was established in 1999 by Lund University to process the exploitation of research results from a research group at the university. The contribution of the company to these projects was in the form of development work, but there were no financial contributions to the university-based research group.

Acknowledgements

This thesis would not have been possible without the efforts of many people, to whom I whish to express my most sincere gratitude. I would like to give special thanks to:

Lars Edenbrandt, my chief supervisor, who gave me great guidance into the world of research and toward this thesis, for always being available when I had questions or problems, for believing in me, and for his never-ending enthusiasm. Thank you for giving me the honour of working with you.

Madis Suurkula, my supervisor, for teaching me the interpretations of bone scans and for being our "Gold".

Iman Hamadeh, Pierre Nordblom, David Jakobsson, Fredrik Olofsson, Andreas Järund, my co-authors, for sharing your knowledge in the field of image processing. Thank you for giving me an insight into the world of engineering.

Mattias Ohlsson, my co-author, for invaluable help with artificial neural networks.

Peter Höglund, my co-author, for helping me with the statistical operation and for fruitful discussions and support.

All participating readers, without you these studies would have been impossible.

My colleagues and friends at the Departments of Nuclear Medicine, Clinical Physiology and Radiophysics for your support and encouragement.

Odd Bech-Hanssen, for opening my eyes to science.

Radomir Uzelac for helping me process the bone scan image files.

Maria Edenbrandt for helping me, at late hours, to convert my figures to Tiff-format.

Jenny Sandgren for your kindness and help with practical matters.

My mother and father, **Lamia and Ali**, my brother **Maher** and his beautiful family, **Katarina, Klara, Elias and David**, my father-in-law, **Hameed**, my friends and relatives scattered all over the world, for supporting and sharing beautiful moments with us.

My two daughters, **Sara and Linn**, for giving me the most wonderful moments in life.

….and of course my husband and my model in life, **Riadh Sadik**, for believing in me, encouraging me and for all the long methodological discussions on our walks contributing with great ideas to this thesis.

References

1. Swayne LC. The private-practice perspective of the manpower crisis in radiology: greener pastures? *J Am Coll Radiol.* 2004; 1:834–841.
2. Berlin L. Liability of interpreting too many radiographs. *AJR Am J Roentgenol.* 2000; 175:17–22.
3. Svensson E, Holm S. Separation of systematic and random differences in ordinal rating scales. *Stat Med.* 1994; 13:2437–2453.
4. Svensson E, Starmark JE, Ekholm S, von Essen C, Johansson A. Analysis of interobserver disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans. *Neurol Res.* 1996; 18:487–494.
5. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph.* 2007; 31:198–211.
6. Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *Am J Roentgenol.* 2008; 190:854–859.
7. Golding JG, Brown MS, Petkovska I. Computer-aided diagnosis in lung nodule assessment. *J Thorac Imaging.* 2008; 23:97–104.
8. Kawamoto, K, Houlihan, CA, Balas, EA, Lobach, DF. Improving clinical practice using clinical decision support systems: a systemic review of trials to identify features critical to success. *BMJ.* 2005; 330:765–768.
9. Doi K. Current status and future potential of computer-aided diagnosis in medical imaging. *Br J Radiol.* 2005; 78:3–19.
10. Freer, TW, Ulissey, MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology.* 2001; 220:781–786.
11. Warren Burhenne LJ, Wood SA, D'orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology.* 2000; 215:554–562.
12. Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. *Radiology.* 2005; 236:451–457.
13. Cupples TE, Cunningham JE, Reynolds JC. Impact of computer-aided detection in a regional screening mammography program. *AJR Am J Roentgenol.* 2005; 185:944–950.

14. Baker ME, Bogoni L, Obuchowski NA, et al. Computer-aided detection of colorectal polyps: can it improve sensitivity off less-experienced readers? Preliminary findings. *Radiology*. 2007; 245:140–149.
15. Taylor SA, Greenhalgh R, Ilangovan R, et al. CT colonography and computer-aided detection: effect of false-positive results on reader specificity and reading efficiency in a low-prevalence screening population. *Radiology.* 2008; 247:133–140.
16. Halligan S, Altman DG, Mallett S, et al. Computed tomographic colonography: assessment of radiologist performance with and without computer-aided detection. *Gastroenterology.* 2006; 131:1690–1699.
17. Mango LJ. Computer-assisted cervical cancer screening using neural networks. *Cancer Lett.* 1994; 77:155–162.
18. Nieminen P, Hakama M, Viikki M, Tarkkanen J, Anttila A. Prospective and randomised public-health trial on neural network-assisted screening for cervical cancer in Finland: results of the first year. *Int J Cancer.* 2003; 103:422–426.
19. Lindahl D, Lanke J, Lundin A, Palmer J, Edenbrandt L. Improved classifications of myocardial bull's-eye scintigrams with computer-based decision support system. *J Nucl Med.* 1999; 40:96–101.
20. Tägil K, Bondouy M, Chaborel JP, et al. A decision support system improves the interpretation of myocardial perfusion imaging. *Eur J Nucl Med Mol Imaging.* 2008; 35:1602–1607.
21. Tryciecky EW, Gottschalk A, Ludema K. Oncologic imaging: interactions of nuclear medicine with CT and MRI using the bone scan as a model. *Semin Nucl Med*. 1997; 27:142–151.
22. Sergieva S, Kirova G, Dudov A. Current diagnostic approaches in tumor-induced bone disease. *J BUON.* 2007; 12:493–504.
23. Abuzallouf S, Dayes I, Lukka H. Baseline staging of newly diagnosed prostate cancer: a summary of the literature. *J Urol.* 2004; 171:2122–2127.
24. Bombardieri E, Gianni L. The choice of the correct imaging modality in breast cancer management. *Eur J Nucl Med Mol Imaging*. 2004; 31:179–186.
25. Myers RE, Johnston M, Pritchard K, Levine M, Oliver T, and the Breast Cancer Disease Site Group of the Cancer Care Ontario Practice Guidelines Initiative. Baseline staging tests in primary breast cancer: a practice guideline. *CMAJ.* 2001; 164:1439–1444.

26. Bombardieri, E, Aktolun, C, Baum, RP, Bishof-Delaloye, A, Buscombe, J, Chatal, JF, et al. Bone scintigraphy procedures guidelines for tumour imaging. *Eur J Nucl Med Mol Imaging.* 2003; 30:107–114.
27. Rossing N, Munck O, Nielsen SP, Andersen KW. What do early bone scans tell about breast cancer patients? *Eur J Cancer Clin Oncol.* 1982; 18:629–636.
28. Peters AM, Bomanji J, Costa DC, Ell PJ, Gordon I, Henderson BL, et al. Clinical audit in nuclear medicine. *Nucl Med Commun*. 2004; 25:97–103.
29. Erdi, YE, Humm, JL, Imbriaco, M, Yeung, H, Larson, SM. Quantitative bone metastases analysis based on image segmentation. *J Nucl Med.* 1997; 38:1401–1406.
30. Noguchi, M, Kikuchi, H, Ishibashi, M, Noda, S. Percentage of positive area of bone metastasis is an independent predictor of disease death in advanced prostate cancer. *Br J Cancer.* 2003; 88:195–201.
31. Yin, T-K, Chiu, N-T. A computer-aided diagnosis for locating abnormalities in bone scintigraphy by a fuzzy system with a three-step minimization approach. *IEEE Trans Med Imaging.* 2004; 23:639–654.
32. Cootes TF, Hill A, Taylor CJ, Haslam J. Use of Active Shape Models for locating structures in medical images. *Image and Vision Computing.* 1994; 12:355–366.
33. Steen PM. Approaches to predictive modeling. *Ann Thorac Surg.* 1994; 58:1836–1840.
34. Cross, SS, Harrison, RF, Kennedy, LR. Introduction to neural networks. *Lancet.* 1995; 346:1075–1079.
35. Rumelhart DE, McClelland JL, eds. *Parallel distributed processing.* Volumes 1&2. Cambridge, MA: MIT Press; 1986.
36. Layfield LJ. Cytologic diagnosis of osseous lesions: A review with emphasis on the diagnosis of primary neoplasms of bone. *Diagn Cytopathol.* 2009; 37:299–310.
37. Sajn L, Kononenko I, Milcinski M. Computerized segmentation and diagnostics of whole-body bone scintigrams. *Comput Med Imaging and Graph*. 2007; 31:531–541.
38. Ellis RL, Meade AA, Mathiason MA, Willison KM, Logan-Young W**.** Evaluation of computer-aided detection systems in the detection of small invasive breast carcinoma. *Radiology.* 2007; 245:88–94.

39. Dorfman DD, Berbaum, KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol*. 1992; 27:723–731.
40. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy B A. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. *Acad Radiol*. 1998; 5:591–602.
41. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. *Acad Radiol*. 2004; 11:1260–1273.
42. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Stat Med.* 2005; 24:1579–1607.
43. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. *Acad Radiol*. 2005; 12:1534–1541.
44. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat Med*. 2007; 26:596–619.
45. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol*. 2008; 15:647–661.
46. Wagner RF, Metz CE, Campell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Acad Radiol.* 2007; 14:723–748.
47. Gilbert FJ, Astley SM, Gillan MG, Agbaje OF, Wallis MG, James J, Boggis CR, Duffy SW; CADET II Group. Single reading with computer-aided detection for screening mammography. *N Engl J Med*. 2008; 359:1675–1684.
48. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, Berns EA, Cutter G, Hendrick RE, Barlow WE, Elmore JG. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med.* 2007; 356:1399–1409.