Lifelong learning in radiology

LIFELONG LEARNING IN RADIOLOGY:
ALL EYES ON VISUAL EXPERTISE

The research reported here was carried out at

in the School of Health Professions Education

in the context of the research school

(Interuniversity Center for Educational Research)

and it was partially funded by a Kootstra Talent Fellowship.

ISBN: 978-94-6421-017-0

Cover design:	Mirakels Ontwerp by Miranda Dood
Layout:	Karin Sanders
Production:	Ipskamp

LIFELONG LEARNING IN RADIOLOGY:
ALL EYES ON VISUAL EXPERTISE

DISSERTATION

To obtain the degree of doctor at Maastricht University,
on the authority of the Rector Magnificus, prof. dr. Rianne M. Letschert in
accordance with the decision of the Board of Deans,
to be defended in public on November 5th 2020, at 16:00 hours.

by

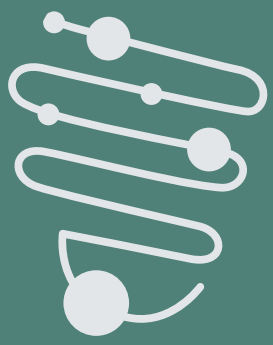
Koos van Geel

Promotors
Prof. dr. J.J.G. van Merriënboer
Prof. dr. S.G.F. Robben

Copromotor
Dr. E.M. Kok

Assessment Committee
Prof. dr. C.P.M. van der Vleuten (chair)
Prof. dr. A.B.H. de Bruin
Dr. A.A. Jacobi
Prof. dr. H. Jarodzka (Open Universiteit Heerlen)
Dr. D.R. Rutgers (UMC Utrecht)

CONTENTS



Chapter 1

General introduction

1

GENERAL INTRODUCTION

Radiology is the medical field concerned with the production and evaluation of medical images (1). Through various imaging techniques, such as conventional radiography, ultrasonography, computed tomography, magnetic resonance imaging, and nuclear medicine, visualizations of the human body are produced. These visualizations contain a wealth of information about patients' anatomy, physiology, and potential pathological processes. Therefore, medical images play a pivotal role in the diagnostic processes of everyday medical practice (1, 2). Furthermore, medical images are increasingly obtained, and new medical imaging techniques are still being developed (3). Additionally, medical images are nowadays readily available with the advent of Picture Archiving and Retrieval Systems (PACS), which enable non-radiology physicians to evaluate images themselves on any computer throughout the hospital (4). The role of medical images in everyday medical practice is thus increasing. However, the evaluation of medical images is not straight-forward (5-7). Abnormalities vary greatly in shape, size, signal characteristics, and enhancement through various contrast agents, while normal anatomical variants may mimic abnormal findings (8, 9). The evaluation of medical images requires the simultaneous processing of many different visual elements. Medical image evaluation is thus considered a complex skill, demanding comprehensive visual knowledge (5, 10). Many years of practice and training are necessary for learning to evaluate medical images (11-13). Improved understanding of how learning to evaluate images takes place could enhance learning.

There are remarkable differences between novices and experts (11), which is also reflected in how they benefit from learning experiences. Novices generally have their first encounters with medical image evaluation in a teaching setting (14). It is thus relevant for novices to investigate how teaching image evaluation can be designed as effective and efficient as possible (15). Intermediates, such as residents in radiology, typically engage in workplace learning. It is thus essential for intermediates to understand how the evaluation process develops over time, as this could provide additional input for feedback to monitor learning (16).

Furthermore, even experts, such as senior radiologists, will never reach a point where learning is completed as they have to adapt to an ever-changing working environment (2, 3). New imaging techniques are frequently introduced, and radiologists with expertise on well-established medical

images may be considered novices on the images produced by new imaging techniques. For radiologists, more insights into optimally implementing these new medical images into their current work field are essential.

This Ph.D. thesis is designed to be of interest to all learners of the expertise spectrum in radiology, from medical students to residents in radiology to senior radiologists. In this general introduction, the theoretical framework of visual expertise development will be introduced. Moreover, eye-tracking methodology, a technique frequently used to study visual expertise, will be presented. Next, the literature on teaching radiology to novices will be summarized, and current gaps in the literature that lead to the first, second, and third research questions are introduced. Subsequently, to study learning in intermediates, a study about the longitudinal development of visual search patterns and lesion detection skills of residents in radiology is presented, which answers the fourth research question. To focus on experts, the use of a novel imaging technique is investigated, which led to the fifth and final research question of this thesis. Finally, to close this general introduction, an outline of the whole Ph.D. thesis will be presented wherein the research questions are coupled to the corresponding studies and Chapters.

Development of visual expertise development in radiology

Laypeople are often amazed about the accuracy and speed of experts in radiology (5, 17), as radiologists can accurately diagnose diseases in mere seconds (18). Investigations about the nature of visual expertise in radiology began as early as the 1940s, yet gained momentum around the 70s (19). One of the first models to describe the nature of visual expertise in radiology was the global-focal search model of Kundel (20), consisting of a global and a focal phase that together constitute the visual search of experts. At first glance, the expert obtains a global impression of the image (18). During this fast and automated global phase, any deviation from normal is noticed (17). During the focal phase, these deviations will subsequently be further inspected and analyzed (5), which will result in a decision whether the deviations are truly manifestations of a particular disease. Most of the currently popular theories on visual expertise still acknowledge some global and focal phase, although more recent models consider both phases to occur parallel or iterative, not necessarily serial (21).

The expertise of the evaluator influences both the global and focal search phases. The expert has seen many thousands of normal cases and has created a mental image of normal (10). This mental image is used in the global search phase to detect any deviations from normal. Novices have not yet created such a mental image of normal and will have to rely solely on focal search (9). Moreover, even if the novice detects an abnormal area, the novice will struggle to diagnose the underlying disease accurately (9). In contrast to the novice, the expert has seen thousands of manifestations of various illnesses. With each abnormality, the expert knowledge basis on radiological manifestations of illnesses has grown and became increasingly organized in meaningful patterns, similar to illness scripts or schemata (22-24). These patterns enable the expert to diagnose diseases during the focal search phase accurately. While a novice may only see an abnormal area, an expert can accurately differentiate whether it most likely represents pneumonia or a tumor (9, 22).

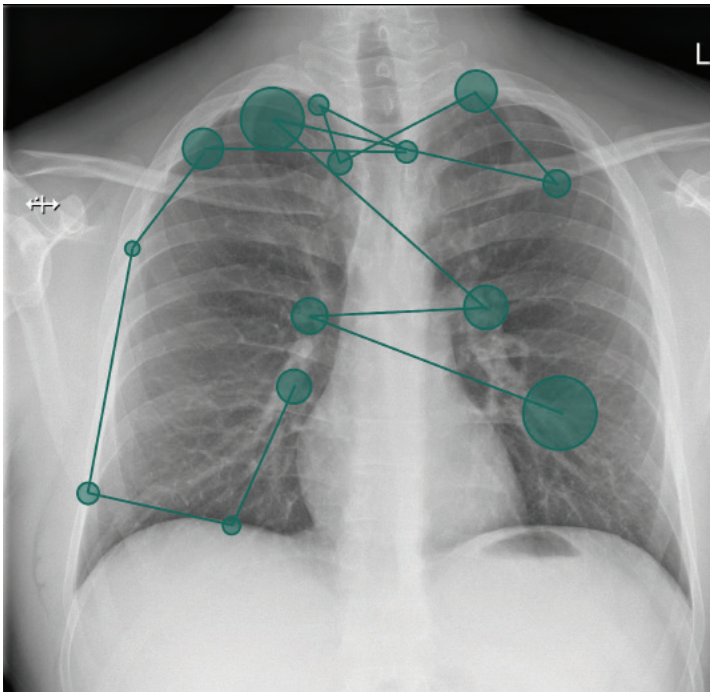
Eye-tracking as a methodology to investigate visual expertise

Differences in visual search on medical images are generally investigated with eye-tracking methodology as follows (25). A participant evaluates an experimental image while the eye tracker records the positions of the eyes. The eye positions constantly change during this evaluation process and can thus provide insights into what is happening during the evaluation process (26). When the eyes stand relatively still, the eyes can process the visual information. Those moments are known as fixations (25). Saccades are the rapid eye movements between fixations, and people are functionally blind during saccades. Particularly the characteristics of fixations and saccades, such as the number of fixations, average fixation duration, time-to-first fixation on an abnormality, and the proportion of abnormal dwell time (the sum of fixations durations on abnormal areas, divided by total sum of fixation durations), and saccade lengths, are commonly used to compare the visual search of novices and experts (25, 27). In Figure 1, the eye movements of a novice evaluating a chest radiograph are visualized.

Based on the visual expertise theories, one can hypothesize differences in eye movement measures of experts compared to novices (27). By their global impression of the medical image, experts can quickly identify abnormalities and divert their visual attention to these abnormalities (21). This global impression generally leads to a lower number of fixations, lower average fixation durations, lower time-to-first fixation on abnormalities, a

higher proportion of abnormal dwell time, and longer saccades of experts compared to novices (27). Indeed, in two meta-analyses on the expertise' development in various domains and on the expertise' development in radiology, evidence was found for the hypothesized differences of experts compared to novices (27, 28).

Figure 1. Visualization of fixations and saccades of a novice evaluating a chest radiograph. Circles visualize fixations, with larger circles indicating longer fixation durations. Lines between circles visualize the saccades.



Nonetheless, it is also essential to acknowledge the influence of (medical) image characteristics on eye movement measures (9, 29, 30). For example, chest radiographs' characteristics are substantially different from mammograms, which will naturally lead to differences in eye movement measures between the two image types. Additionally, some parameters are even of little use because of image characteristics. For example, when a disease affects the whole lungs, such as lung fibrosis, it is of little value to calculate the time-to-first fixation because it will be virtually zero in every case (9). Thus, the influence of image characteristics should always be taken into account when using eye-tracking technology.

Current gaps in teaching radiology

While studies on the development of visual expertise can provide valuable information about how learning radiology over the whole expertise spectrum takes place, they do not provide clear-cut answers on how to design radiology teaching initiatives for novices and non-radiology physicians. In recent decades many radiology teaching initiatives have been initiated. More and more non-radiology physicians started to evaluate medical images because they have become readily available by PACS (4, 31). Nowadays, it is not uncommon that diagnostic decisions and treatment plans are made before an image has been evaluated by a radiologist (32, 33). Non-radiology physicians, particularly inexperienced physicians, make significantly more errors in image evaluations than radiologists (34-36). Since non-radiology physicians are somehow expected to evaluate images themselves, one should ask how they are prepared for this task. The amount of training in image evaluation varies but can be as low as one hour for the whole undergraduate medical curriculum (37). Many medical students do not feel they are properly trained to evaluate images independently (38). Teachers may struggle with the question of how to prepare medical students to evaluate medical images optimally. With medical curricula already saturated and little room for more training (32), radiology teaching initiatives must provide the most effective and efficient learning experiences (15). Some studies have investigated how to provide effective and efficient learning experiences in image evaluation training (15). These studies typically have an (implicit) theoretical background based on the theories of illness scripts or visual expertise.

Based on theories of illness scripts, the studies on medical image evaluation training have primarily focused on *how to analyze abnormal findings*. The development of illness scripts requires an extensive knowledge basis of normal anatomy and normal variants, pathophysiological processes of various diseases, and characteristics of the patient population, such as the prevalence of diseases (22). Disease prevalence can influence the criterion of physicians to call a finding abnormal or normal (39). There is currently a substantial discrepancy between the prevalence of diseases in image evaluation training and medical practice: image evaluation training generally focuses on abnormal findings (40), whereas the majority of medical images in clinical practice -- for example, on a ward or an emergency department -- are normal (41, 42). The prevalence of normal and abnormal cases in a radiology teaching initiative could impact the criterion of less experienced evaluators.

1

Indeed, a sensitivity-specificity tradeoff based on the prevalence effect was found on sensitivity (correctly identified abnormalities) and specificity (correctly identified images without abnormalities) of residents evaluating ankle radiographs by Pusic et al. (40): The residents who predominantly evaluated radiographs with ankle fractures became more sensitive to detect ankle fractures, whereas residents who predominantly evaluated radiographs without ankle fractures developed to be more specific. These participants had some clinical experience as residents and may already have developed a criterion based on the prevalence in their own practice. Inexperienced medical students may be influenced particularly by a wrong prevalence in image evaluation training since they have not been exposed to the prevalence of clinical practice yet. The first research question is thus as follows:

> *RQ1 What are the effects of the prevalence of normal images in a practice phase of medical image evaluation training on third-year medical students` detection and analysis of abnormalities?*

Moreover, some studies have investigated the instructional design of medical image evaluation training. Educational designs generally consist of an explanation by an (expert) teacher and practice by the learner (14). When to provide an expert's explanation remains a debate in medical education (43). Explanation prior to practice, also known as a deductive instructional sequence, has the advantage that it is time-efficient. Deductive instructional sequences are advocated for learners who already have some prior knowledge on the subject (14). Practice prior to explanation, also known as an inductive instructional sequence, has the advantage that students have to figure out solutions for themselves (44). They may not always find this solution but become immersed in the problem, known as productive failure. Productive failure should lead to a deeper understanding of the problem (44). Medical students could benefit from inductive instructional sequences in particular since image evaluation training is generally scarce in medical schools (37), and students have little knowledge of the subject.

However, the differences between inductive and deductive learning on medical image evaluation training have not yet been investigated, which has led to the second research question:

RQ2 What are the effects of an inductive and deductive instructional sequence in medical image evaluation training on third-year medical students` detection and analysis of abnormalities?

Moreover, in order to analyze abnormalities, one first has to search for abnormalities. The topic of *how to search for abnormal findings* is thus also frequently covered in image evaluation training. Novices are generally instructed to adopt a systematic approach. A systematic approach refers to visual searches that are always performed in the same specific order (45). The rationale behind a systematic approach is that novices, as they learn to repeatedly use the same evaluation order, do not forget to evaluate areas of a medical image, ultimately leading to complete evaluations and fewer missed abnormalities (1, 37, 45, 46).

A few studies have investigated the effects of systematic viewing training on visual search patterns and the detection of abnormalities (45, 47, 48). Positive effects of a systematic viewing training on chest nodule detection by medical trainees were found by Auffermann et al. (47). However, this investigation lacked a control condition with an identical expert's explanation of chest radiograph evaluations. Thus, it is unknown whether the increased detection of chest nodules was caused by teaching a systematic viewing strategy, or the expert's explanation. When compared to a nonsystematic control condition, no effects of a systematic viewing training were found on the detection of abnormalities on chest radiographs by third-year medical students by Kok et al. (45). The students of the systematic search group became more systematic and had more complete visual searches, yet this did not result in increased detection. These third-year medical students had to search for various abnormal findings, ranging from signs indicating heart failure to rib fractures (45). These participants may have been too inexperienced to oversee all of the abnormal findings as they were only third-year medical students. Perhaps these students were thus not capable yet of adopting a systematic approach for the complete range of abnormal findings of this study. (Future) physicians are expected to evaluate common abnormal findings in their future clinical work, and it is thus essential to investigate the effects of systematic viewing training on more experienced novices, such as final-year medical students. Thus, the third research question is as follows:

RQ3 Do visual search patterns change, and does the detection of abnormalities increase after systematic viewing training of final-year medical students when evaluating chest radiographs?

1

Mapping expertise development in intermediates

Intermediates typically engage in workplace learning through feedback on their image evaluations. Their learning experience can thus be enriched by direct and rich sources of feedback (16). To provide such feedback, it is essential to improve the understanding of how evaluation processes develop over time. So what is known about this development? It is still challenging to make statements about the longitudinal development of learners in radiology based on the current literature (29, 49, 50). First, the novices in previous research were generally lay-people or medical students, while the experts were radiologists with decades of experience (10). These expertise groups are vastly different. The interpolation of findings between these two diverse groups can thus be troublesome. Second, development is generally not gradual and can have periods of stagnation and temporary decreases, perhaps when knowledge is reorganized (13, 22, 24). Fortunately, some studies have used groups of intermediates, for example, residents in radiology (9).

When intermediates are compared to novices and experts, the picture becomes less black and white. Average fixation durations and time-to-first fixation on abnormalities were different from the hypotheses (27, 51): Based on the visual expertise theories, the intermediates were expected to have intermediate average fixation durations and time-to-first fixations, somewhere in-between the experts and novices (27). However, it was found that the intermediates had the longest average fixation durations and time-to-first fixations, while experts and novices had shorter average fixation durations and time-to-first fixations on abnormalities. The development of visual search in radiology may thus also show periods of stagnation or decreases in performance. A more fine-grained, prospective, longitudinal study on the development of visual search in radiology is yet lacking in the literature (26, 29, 49), and is necessary to enhance feedback and monitoring of intermediate learners. To improve our understanding of the longitudinal development of evaluation processes in intermediates, the fourth research question of this Ph.D. thesis is as follows:

RQ4 How does the longitudinal development of visual search patterns and lesion detection skills of first-year residents in radiology take place when evaluating chest radiographs?

How to deal with new imaging techniques?
After evaluating an almost inestimable number of medical images and with thousands of hours of work experience, radiologists are considered experts in their subspecialties (11, 13, 52). However, this does not mean that the learning process of radiologists will eventually finish. New imaging techniques are still developed, and the medical practice of radiologists keeps evolving. New imaging techniques can be of great benefit to patient care as they enable more accurate and more efficient diagnosis, yet they can raise questions on how to incorporate these new techniques into radiologists` medical practice.

The contrast-enhanced mammogram (CEM) is a technique that has recently been developed and has aided experts in breast radiology in the detection of breast cancer (53-55). For decades plain mammograms have been the worldwide standard for the screening and initial detection of breast cancer (13). However, mammograms are difficult to evaluate when women have dense breast tissue as the glandular tissue may mimic the appearance of malignant tissue (56). An iodine intravenous contrast agent was found to attenuate in malignant tissue and not in benign, glandular tissue. CEM utilizes this difference in attenuation between malignant and benign tissue. After intravenous injection of the iodine contrast agent, two images are produced. First, a low-energy (LE) image is produced, which is similar to a plain mammogram, also known as a full-field digital mammogram (FFDM) (57). Subsequently, an additional, post-contrast image is produced, known as a recombined contrast-enhanced (RC) image (54, 58). Both images are evaluated conjointly and comprise the CEM examination (54). The sensitivity and specificity of breast radiologists for breast cancer detection increase significantly with the addition of this contrast-enhanced RC mammogram to the evaluation (54, 55, 59). Although the RC image increases radiologists' accuracy, it adds an extra image to the evaluation process and increases the workload.

How should radiologists thus deal with this new technique in their medical practice? It is yet unknown how the addition of this new, contrast-enhanced image affects the evaluation process and time necessary to complete

1

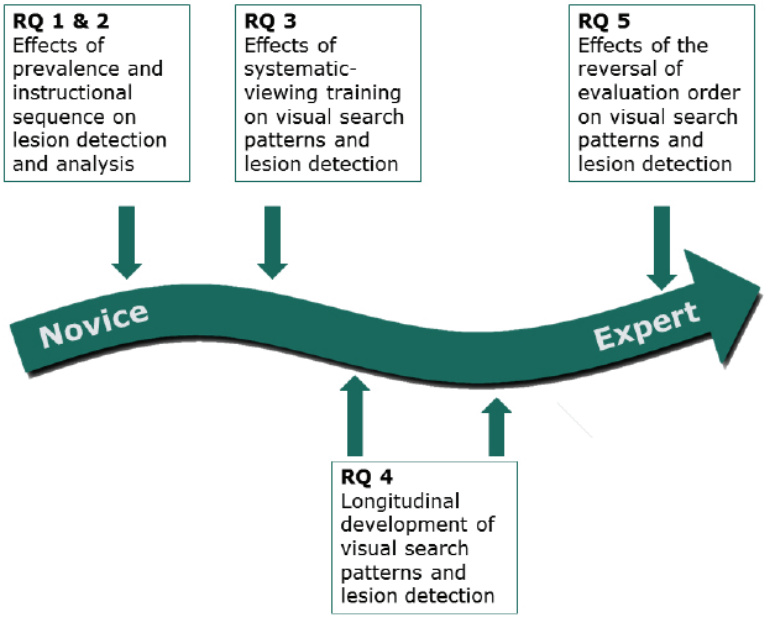
an evaluation. Furthermore, the new contrast-enhanced image could be evaluated before, instead of after, the LE image. Manufacturers currently advise to use an evaluation order wherein the contrast-enhanced image is evaluated after the LE image (53). However, breast radiologists with some experience in evaluating CEM images anecdotally mention evaluating the RC image first since malignant lesions are far more salient on the contrast-enhanced images. It is yet unknown how the evaluation order affects visual search and detection of malignant lesions by breast radiologists, and the fifth research question is thus as follows:

RQ5 What are the effects of reversal of the evaluation order of plain (LE) and contrast-enhanced (RC) mammograms on visual search patterns and the detection of malignant breast lesions by breast radiologists?

Thesis outline

This Ph.D. thesis investigates how learning to evaluate medical images over the range of learners from medical students to radiologists takes place, which can be used to support lifelong learning experiences. See Figure 2 for an overview of the separate yet interconnected Ph.D. studies.

Figure 2. Overview of the Ph.D. dissertation.



First, different strategies to strengthen image evaluation training for novices in radiology are investigated in Chapters 2 and 3 of this thesis: In Chapter 2, an experimental study on the effects of prevalence of normal images in a practice phase and the effects of an inductive versus a deductive instructional design on third-year medical students` detection and analysis of abnormalities is presented. This study answers research questions 1 and 2. Furthermore, in Chapter 3, an experimental eye-tracking study on the effects of systematic viewing training on visual search patterns and detection of abnormalities by final-year medical students is presented, which addresses research question 3.

Second, how learning radiology in intermediates takes place is investigated with a longitudinal, prospective eye-tracking study on the development of visual search patterns and the detection of abnormalities on chest radiographs of first-year residents in radiology. This study is presented in Chapter 4 of this thesis and addresses research question 4.

Third, how to support experts is addressed in Chapter 5 of this thesis. In this chapter, an experimental eye-tracking study on the effects of the evaluation order (LE-RC order versus RC-LE order mammogram followed by plain mammogram) on the evaluation process and detection of malignant breast lesions is presented. This study addresses research question 5.

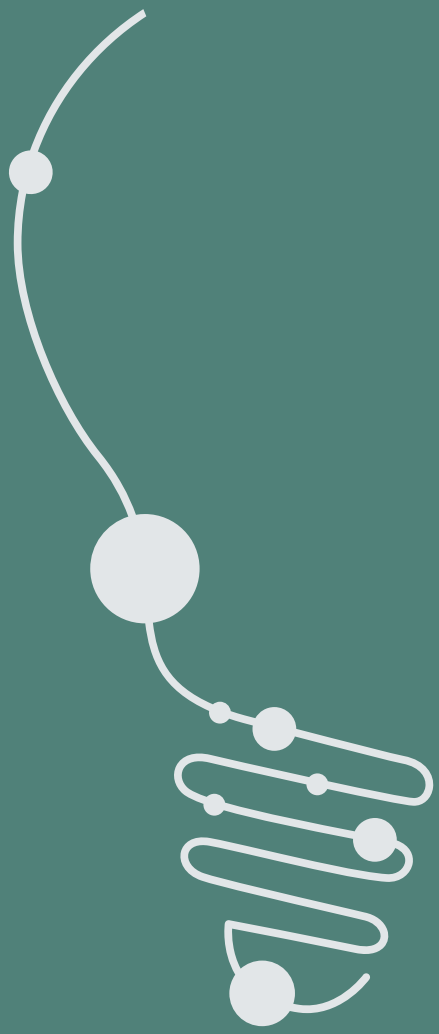
Finally, in the general discussion presented in Chapter 6, the main findings of the separate studies are summarized, and the theoretical and practical values of the combined studies and the limitations are appraised. The general discussion is completed by the thesis` general conclusions on how learning radiology from the range of novices to experts takes place and how lifelong learning in radiology can be supported.

REFERENCES

1. Adam A, Dixon AK, Gillard JH, Schaefer-Prokop C, Allison DJ, Grainger RG. Grainger & Allison's diagnostic radiology : a textbook of medical imaging. 2015.
2. Iglehart JK. The new era of medical imaging--progress and pitfalls. The New England journal of medicine. 2006;354(26):2822-8.
3. Levin DC, Rao VM, Parker L, Frangos AJ. Analysis of radiologists' imaging workload trends by place of service. Journal of the American College of Radiology : JACR. 2013;10(10):760-3.
4. Reiner BI, Siegel EL, Hooper F, Protopapas Z. Impact of filmless imaging on the frequency of clinician review of radiology images. Journal of digital imaging. 1998;11(3 Suppl 1):149-50.
5. Samei E, Krupinski EA. The Handbook of Medical Image Perception and Techniques: Cambridge University Press; 2010.
6. Krupinski EA. Current perspectives in medical image perception. Attention Perception & Psychophysics. 2010;72(5):1205-17.
7. Manning D. Medical Image perception: its achievements, challenges and a role in medical education. Keynote presentation MIPS XIV Dublin, August 2011. 2011.
8. Keats TE, Anderson MW. Atlas of normal roentgen variants that may simulate disease. 2013.
9. Kok EM, de Bruin ABH, Robben SGF, van Merriënboer JJG. Looking in the Same Manner but Seeing it Differently: Bottom-up and Expertise Effects in Radiology. Applied Cognitive Psychology. 2012;26(6):854-62.
10. van der Gijp A, Ravesloot CJ, Jarodzka H, van der Schaaf MF, van der Schaaf IC, van Schaik JPJ, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education. 2017;22(3):765-87.
11. Ericsson KA, Charness N, Feltovich P, Hoffman RR. The Cambridge Handbook of Expertise And Expert Performance: Cambridge University Press; 2006.
12. Ravesloot CJ, van der Schaaf MF, Kruitwagen C, van der Gijp A, Rutgers DR, Haaring C, et al. Predictors of Knowledge and Image Interpretation Skill Development in Radiology Residents. Radiology. 2017;284(3):758-65.
13. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. 2009;253(3):632-40.
14. van Merriënboer JJG, Kirschner PA. Ten Steps to Complex Learning: A Systematic Approach to Four-component Instructional Design: Routledge; 2018.
15. Kok EM, van Geel K, van Merriënboer JJG, Robben SGF. What We Do and Do Not Know about Teaching Medical Image Interpretation. Frontiers in Psychology. 2017;8(309).
16. Kilminster S, Cottrell D, Grant J, Jolly B. AMEE Guide N° 27: Effective educational and clinical supervision. Medical teacher. 2007;29:2-19.
17. Drew T, Evans K, Vo MLH, Jacobson FL, Wolfe JM. Informatics in Radiology What Can You See in a Single Glance and How Might This Guide Visual Search in Medical Images? Radiographics. 2013;33(1):263-74.
18. Kundel HL, Nodine CF. A Visual Concept Shapes Image Perception. Radiology. 1983;146(2):363-8.
19. Kundel HL. History of Research in Medical image perception. JACR. 2006.
20. Kundel HL, La Follette PS, Jr. Visual search patterns and experience with radiological images. Radiology. 1972;103(3):523-8.
21. Reingold EM, Sheridan H. Èye movements and visual expertise in chess and medicine. In: Leversedge SP, Gilchrist ID, Everling S, editors. Oxford Handbook on Eye Movements. Oxford: Oxford University Press; 2011. p. 528-50.
22. Custers EJ. Thirty years of illness scripts: Theoretical origins and practical applications. Med Teach. 2015;37(5):457-62.
23. Barrows HS, Feltovich PJ. The clinical reasoning process. Med Educ. 1987;21(2):86-91.
24. Custers E. Medical Education and Cognitive Continuum Theory: An Alternative Perspective on Medical Problem Solving and Clinical Reasoning. Academic Medicine. 2013;88(8):1074-80.
25. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Weijer J. Eye Tracking: A Comprehensive Guide to Methods

and Measures: Oxford University Press; 2011.
26. Kok EM, Jarodzka H. Before your very eyes: the value and limitations of eye tracking in medical education. Med Educ. 2017;51(1):114-22.
27. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. Educational Psychology Review. 2011:1-30.
28. Gijp A, Schaaf MF, Schaaf IC, Huige JCBM, Ravesloot CJ, Schaik JPJ, et al. Interpretation of radiological images: towards a framework of knowledge and skills. Advances in Health Sciences Education. 2014:1-16.
29. Jarodzka H, Holmqvist K, Gruber H. Eye tracking in Educational Science: Theoretical frameworks and research agendas. Journal of Eye Movement Research. 2017;10(1).
30. Bertram R, Helle L, Kaakinen JK, Svedstrom E. The Effect of Expertise on Eye Movement Behaviour in Medical Image Perception. PloS one. 2013;8(6).
31. Rankin RN. Technologies for Teaching: exploring the use of PACS, databases and Teaching files. In: Chhem RK, Hibbert KM, van Deven T, editors. Radiology Education The scholarship of Teaching and learning. Berlin Heidelberg: Springer-Verlag; 2009.
32. Zwaan L, Kok E, van der Gijp A. Radiology education: a radiology curriculum for all medical students? Diagnosis. 2017;4.
33. Saha A, Roland RA, Hartman MS, Daffner RH. Radiology medical student education: an outcome-based survey of PGY-1 residents. Academic radiology. 2013;20(3):284-9.
43. Mylopoulos M, Brydges R, Woods NN, Manzone J, Schwartz DL. Preparation for future learning: a missing competency in health professions education? Med Educ. 2016;50(1):115-23.
44. Kapur M. Productive Failure. Cognition and Instruction. 2008;26(3):379-425.
45. Kok E, Jarodzka H, Bruin AB, Bin Amir H, Robben SGF, Merrienboer JG. Systematic viewing in radiology: seeing more, missing less? Adv Health Sci Educ Theory Pract. 2015.
46. Berlin L. Radiologic errors and malpractice: a blurry distinction. AJR Am J Roentgenol. 2007;189(3):517-22.
47. Auffermann WF, Little BP, Tridandapani S. Teaching search patterns to medical trainees in an educational laboratory to improve perception of pulmonary nodules. JMIOBU. 2015;3(1):011006-.
48. Auffermann WF, Henry TS, Little BP, Tigges S, Tridandapani S. Simulation for Teaching and Assessment of Nodule Perception on Chest Radiography in Nonradiology Health Care Trainees. Journal of the American College of Radiology : JACR. 2015;12(11):1215-22.
49. Gegenfurtner A, Kok E, van Geel K, de Bruin A, Jarodzka H, Szulewski A, et al. The challenges of studying visual expertise in medical image diagnosis. Med Educ. 2017;51(1):97-104.
50. Kok EM. Eye tracking: the silver bullet of competency assessment in medical image interpretation? Perspectives on medical education. 2019;8(2):63-4.
51. Bertram R, Kaakinen J, Bensch F, Helle L, Lantto E, Niemi P, et al. Eye Movements of Radiologists Reflect Expertise in CT Study Interpretation: A Potential Tool to Measure Resident Development. Radiology. 2016;281(3):805-15.
52. Norman G, Eva KW, Brooks LR, Hamstra S. Expertise in Medicine and Surgery. In: Ericsson KA, Charness N, Feltovich P, Hoffman RR, editors. The Cambridge handbook of expertise and expert performance. Cambridge: Cambridge University Press; 2006. p. 339-53.
53. Bhimani C, Matta D, Roth RG, Liao L, Tinney E, Brill K, et al. Contrast-enhanced Spectral Mammography: Technique, Indications, and Clinical Applications. Academic radiology. 2017;24(1):84-8.
54. Patel BK, Lobbes MBI, Lewin J. Contrast Enhanced Spectral Mammography: A Review. Seminars in ultrasound, CT, and MR. 2018;39(1):70-9.
55. Tagliafico AS, Bignotti B, Rossi F, Signori A, Sormani MP, Valdora F, et al. Diagnostic performance of contrast-enhanced spectral mammography: Systematic review and meta-analysis. Breast (Edinburgh, Scotland). 2016;28:13-9.
56. Sorin V, Yagil Y, Yosepovich A, Shalmon A, Gotlieb M, Neiman OH, et al. Contrast-Enhanced Spectral Mammography in Women With Intermediate Breast Cancer Risk and Dense Breasts. AJR Am J Roentgenol. 2018:W1-w8.

57. Lalji UC, Jeukens CR, Houben I, Nelemans PJ, van Engen RE, van Wylick E, et al. Evaluation of low-energy contrast-enhanced spectral mammography images by comparing them to full-field digital mammography using EUREF image quality criteria. European radiology. 2015;25(10):2813-20.
58. Lalji UC, Houben IP, Prevos R, Gommers S, van Goethem M, Vanwetswinkel S, et al. Contrast-enhanced spectral mammography in recalls from the Dutch breast cancer screening program: validation of results in a large multireader, multicase study. European radiology. 2016;26(12):4371-9.
59. Houben IPL, Van de Voorde P, Jeukens C, Wildberger JE, Kooreman LF, Smidt ML, et al. Contrast-enhanced spectral mammography as work-up tool in patients recalled from breast cancer screening has low risks and might hold clinical benefits. Eur J Radiol. 2017;94:31-7.



Chapter 2

Chest X-ray evaluation training: impact of normal and abnormal image ratio and instructional sequence

Koos van Geel, Ellen M. Kok, Abdullah D. Aldekhayel, Simon G.F. Robben, Jeroen J.G. van Merriënboer.

Medical Education. 2019;53(2):153-64.

2

ABSTRACT

Context

Medical image perception training generally focuses on abnormalities, whereas normal images are more prevalent in medical practice. Furthermore, instructional sequences that let students practice prior to expert instruction (inductive) may lead to improved performance compared with methods that give students expert instruction before practice (deductive). This study investigates the effects of the proportion of normal images and practice-instruction order on learning to interpret medical images. It is hypothesized that manipulation of the proportion of normal images will lead to a sensitivity specificity trade-off and that students in practice-first (inductive) conditions need more time per practice case but will correctly identify more test cases.

Methods

Third-year medical students (n = 103) learned radiograph interpretation by practicing cases with, respectively, 30% or 70% normal radiographs prior to expert instruction (practice-first order) or after expert instruction (instruction-first order). After training, students made a test (60% normal), and sensitivity (% correctly identified abnormal radiographs), specificity (% correctly identified normal radiographs), diagnostic performance (% correct diagnoses), and case duration were measured.

Results

The conditions with 30% normal images scored higher on sensitivity, but the conditions with 70% normal images scored higher on specificity, indicating a sensitivity and specificity trade-off. Those who participated in inductive conditions took less time per practice case but more per test case. They had similar test sensitivity but scored lower on test specificity.

Conclusions

The proportion of normal images impacted the sensitivity-specificity trade-off. This trade-off should be an important consideration for the alignment of training with future practice. Furthermore, the deductive conditions unexpectedly scored higher on specificity, while participants took less time per case. An inductive approach did not lead to higher diagnostic performance, possibly because participants might already have relevant prior knowledge. Deductive approaches are therefore advised for the training of advanced learners.

INTRODUCTION

The interpretation of medical images, such as electrocardiograms, pathology slices, or radiographs, is an important part of everyday medical practice (1-3). Research on medical image interpretation has primarily focused on the characteristics of visual expertise (3, 4). In such research, novices and experts in image interpretation are compared, and the experts` performance is superior to that of novices. Experts also show more efficient viewing behavior (5). Although such research on visual expertise provides invaluable information on how learning to interpret images takes place, it does not provide straightforward answers regarding questions to teaching medical image perception. The current study aims to add to the literature regarding: (i) the "what" content of medical image perception training; and the "how" instructional design of medical image perception training.

The content of medical image perception training

Concerning the content of image interpretation training, there is generally a large emphasis on abnormal images (2). Only a small amount of time in medical curricula is devoted to teaching image interpretation (6), whereas a vast amount of anatomy and (patho)physiology needs to be covered. Although it might be time-efficient, this emphasis on abnormal images may also give students a wrong impression about the prevalence of diseases in medical practice. In reality, many images in everyday clinical practice in a ward or an emergency department are found to be normal or do not contain significant or relevant pathology (7-9). This mismatch between the low prevalence of diseases in clinical practice and the emphasis on abnormal images during training can impact students` performance in practice. Indeed, Pusic et al. (2) have shown that a change of the proportion of abnormal practice cases alters the sensitivity (proportion of correctly identified abnormal images out of the total number of abnormal images) and specificity (proportion of correctly identified normal images out of the total number of normal images) of the performance of emergency residents. The residents who practiced with predominantly abnormal images had higher sensitivity, whereas the residents who practiced with predominantly normal images had higher specificity. The emergency residents in the study by Pusic et al. (2) already had some experience interpreting medical images and might have learned about the low prevalence of diseases in clinical practice. To what extent medical students are impacted by the proportion of normal images in training is not yet known. It is expected that the performance of more novice students potentially increases even more

when they are trained with a high proportion of normal images in medical image perception training.

Instructional design of medical image perception training

The instructional design of image interpretation training, like most educational experiences, often consists of a presentation by an expert, the practice of the task by learners, and feedback. When to provide expert instruction and practice for an effective educational experience remains a debate in medical education (10). Direct or deductive-expository instruction, which starts with the expert instruction followed by a practice phase, is advocated for more advanced learners, when instructional time is limited, and when a deep level of understanding is not strictly necessary (11).

By contrast, inductive approaches such as problem-based learning and guided discovery learning (12) offer practice prior to instruction. As students first practice, they will have to figure out solutions for themselves instead of only implementing a solution presented by an expert. Students may fail to find the solution and will need more time to complete a practice case. However, this failure may be considered productive (13). Students are fully immersed in the problem when searching for the solution. This productive failure can, therefore, lead to a deeper understanding and long-term retention of knowledge (14). The benefits of productive failure indeed have been shown in research in mathematics education (15). Despite the theoretical benefit of inductive approaches, most medical image interpretation training still use deductive approaches. It is therefore not known if productive failure can be induced in medical students who are learning to interpret medical images.

The present study

In this study, the effects of the proportion of normal images (30% versus 70% normal) and instructional sequence (deductive versus inductive) in a chest radiograph perception training on the performance of third-year medical students were investigated.

Research questions

1. What are the effects of the proportion of normal images in a practice phase of medical image perception training on third-year medical students` performance?

2. What are the effects of instructional sequencing (inductive or deductive) in image perception training on third-year medical students` performance?

The students` performance was defined as sensitivity, specificity, diagnostic performance, and case duration on a subsequent test.

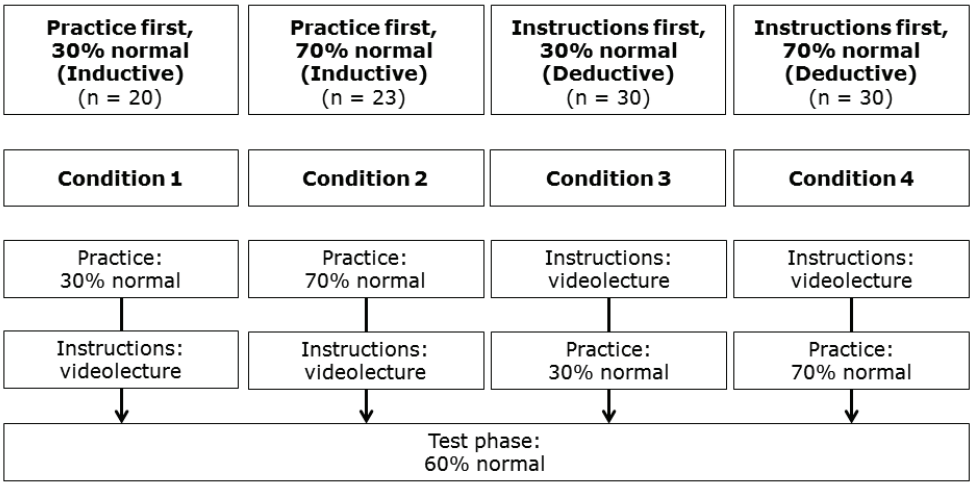
Hypotheses
In line with Pusic, we hypothesize that on a post-test:
1. Students practicing with a low proportion of normal images will have higher sensitivity scores, whereas students practicing with a high proportion of normal images will have higher specificity scores.
2. Students in inductive conditions will have higher sensitivity, specificity, and diagnostic performance than students in deductive conditions.

Concerning students` performance during the practice phase, students in the inductive conditions will be engaged in the act of productive failure, we hypothesize that this should result in:
3. Lower sensitivity, specificity, and diagnostic performance in the inductive conditions.
4. Image interpretation during the practice phase will take more time.
5. Students will need more time per case on those they misinterpret compared with cases they correctly interpret, which will reflect productive failure. This difference will be higher for students in inductive conditions.

METHODS
This 2x2 design tested the effects of proportion (practicing with a proportion of 30% normal images [condition 1 and 3] versus a proportion of 70% normal images [condition 2 and 4]) and instructional sequence (a practice-first [inductive] [conditions 1 and 2] versus an instruction-first [deductive] sequence [conditions 3 and 4]) (Fig. 1). After the training, students` sensitivity, specificity, diagnostic performance, and case duration were measured in a test with a proportion of normal images typical of everyday clinical practice.

Participants
A total of 103 third-year medical students took part in this study (69% female; mean age = 22.5 ± 2.43 years) from Maastricht University in the Netherlands. All students were approached via announcements prior

Figure 1. Flowchart of the 2 × 2 design on the four experimental conditions.



to regular lectures and via announcements on the electronic learning environment of Maastricht University in September 2016. None of the participants had yet received any formal training in interpreting chest radiographs. Participants were randomly assigned to the four experimental conditions in a 2x2 design (Fig. 1).

Two of the conditions started with the practice phase, consisting of practicing with 20 chest radiographs. The proportion of normal radiographs during the practice phase was manipulated (70% normal radiographs versus 30% normal radiographs). The other two conditions started with the instructions phase, consisting of a video lecture, and subsequently practiced with a set with either 70% normal or 30% normal images, yielding a full 2x2 design. The participants received a €20 gift voucher after the experiment as compensation. All participants signed informed consent, and the ethical review board of the Dutch Association for Medical Education (NVMO-ERB) approved this study, file number 763.

MATERIALS

Video lecture

During the instruction phase, a video lecture was used. This video lecture was designed for this experiment by AA and SGFR. The video covered the basics of chest radiograph interpretation and the radiologic manifestations of eight common abnormalities: pneumonia, pneumothorax, pleural effusion,

atelectasis, lung tumors, cardiomegaly, emphysema, and bilateral hilar lymphadenopathy. Two normal chest radiographs and two examples of each abnormality were used in the video, which totaled 18 radiographs. The video had a duration of 23 minutes, and participants saw the video only once. Participants were not allowed to stop, rewind, or fast-forward the video. Furthermore, participants were not allowed to make notes. The video lecture was shown individually to participants using Windows Movie Player 12 (Microsoft Corp., Redmond, WA, USA).

Radiological images

The radiographs used in this experiment originated from a teaching file consisting of over 400 chest radiographs from the radiology department of Maastricht University Medical Center. All radiographs were stripped of any patient information and were selected by KG and SGFR. Radiographs were selected to have no abnormalities (thus normal images) or only one type of the eight previously mentioned cardiopulmonary pathologies. The cases have been used in previous investigations involving third-year medical students as well as final-year medical students (16, 17). In the investigation with final-year medical students, learning effects were visible after practicing with 10 cases and a video lecture. To ensure a learning effect in third-year medical students, the number of cases in the practice phase was doubled. The images used in this investigation are available upon request from the first author (KG).

In the practice phase, participants interpreted 20 chest radiographs; 12 radiographs were identical in each of the four conditions, with half of these identical radiographs being normal. In conditions 1 and 3, for the 30% normal images, the other eight radiographs were abnormal. In conditions 2 and 4, for the 70% normal images, the other eight chest radiographs were normal. See Figure 2 for an example of a chest radiograph used in the experiment.

The test phase consisted of 20 chest radiographs, of which 60% were normal images. In daily practice, normal images predominate over abnormal images (7). The order of the 20 radiographs was randomized per participant. The cases were heterogeneous in variance: abnormal cases, ($F(7, 102) = 13.4$, $p < .001$), normal cases: ($F(11, 102) = 35.4$, $p < .001$), diagnostic performance: ($F(7, 102) = 25.3$, $p < .001$).

Calculation of Cronbach's alpha would produce unreliable estimates (18). Instead, Macdonalds` Ω_t was calculated, which can be similarly interpreted to a Cronbach's alpha. The Ω_t of abnormal cases was .63, the Ω_t of normal cases was .67, and the Ω_t of diagnostic performance was .41. The characteristics of the test phase (discrimination of the cases, mean percentage correctly identified, average case duration, and diagnosis per case) can be found in Table 1.

Figure 2. Example of a chest x-ray used in the experiment.

Table 1. Characteristics of the test phase; Case discriminations, mean percentage correctly identified, average case duration and diagnosis per case.

Case	Discrimination	Mean % correctly discriminated (*SD*)	Case duration in seconds (*SD*)	Diagnosis
1	Normal	51 (50)	34 (17)	
2	Abnormal	99 (1)	28 (15)	Pneumothorax
3	Normal	50 (50)	35 (18)	
4	Normal	82 (39)	28 (17)	
5	Abnormal	100 (/)	28 (15)	Atelectasis
6	Normal	84 (36)	33 (20)	
7	Abnormal	96 (19)	31 (18)	Cardiomegaly
8	Normal	73 (45)	32 (17)	
9	Normal	72 (45)	32 (17)	
10	Abnormal	94 (24)	32 (19)	Hilar enlargement
11	Abnormal	83 (38)	30 (15)	Lung tumor
12	Normal	18 (39)	41 (19)	
13	Abnormal	100 (/)	38 (17)	Pneumonia
14	Normal	92 (27)	28 (14)	
15	Abnormal	77 (43)	31 (17)	Emphysema
16	Normal	80 (41)	32 (18)	
17	Normal	25 (44)	33 (17)	
18	Normal	79 (41)	29 (16)	
19	Abnormal	96 (14)	27 (16)	Pleural effusion
20	Normal	51 (50)	37 (19)	

MEASURES

Sensitivity and specificity

Sensitivity in the practice phase and test phase was defined as the proportion of abnormal radiographs correctly identified as abnormal. Specificity in the practice phase and test phase was defined as the proportion of normal radiographs correctly identified as normal.

Diagnostic performance

If the participants deemed a radiograph abnormal during the practice phase or test phase, they were requested to type their most probable diagnosis via a free text form. A coding scheme for correct diagnoses and their respective synonyms was developed by KG and SGFR. All correctly diagnosed radiographs were subsequently coded as 1, all incorrect answers were coded as 0. To calculate the diagnostic performance of participants, all diagnosis scores were summed and divided by the total number of eight abnormal radiographs. For the diagnosis scores of the practice phase, only the six abnormal cases that were identical in all four conditions were used.

Average case duration

The time needed by participants to interpret a radiograph and provide answers was registered and averaged for the 12 (normal and abnormal) identical radiographs in the practice phase and the 20 radiographs (cases) in the test phase.

PROCEDURE

The experiment was conducted in 11 sessions, with a maximum of 10 students per experimental session. Every participant worked on a desktop computer with a 22” LCD (liquid crystal display) screen with a resolution of 1650 x 1080 pixels by use of the Qualtrics software (Qualtrics, Provo, Utah, USA)(19). Each session started with a short briefing of five minutes in which the procedure was delineated, and participants subsequently provided written consent. Participants were not informed about the proportion of normal images of the practice phase or test phase. The order of the cases in the test phase was randomized per participant by the QUALTRICS software. Participants worked individually throughout the whole experiment.

During the practice phase, participants had a maximum of 80 seconds to interpret each of the 20 chest radiographs and to report if they were normal or abnormal images. If the image was abnormal, they were required to

report the most probable diagnosis. The time limit of 80 seconds was based on a previous investigation with third-year medical students who took an average of 52.6 seconds (standard deviation [SD] 20.6) to interpret a case (16). Based on these numbers, the probability of not completing a case within 80 seconds would be .09, which was considered acceptable. After 80 seconds, a new page was automatically loaded that informed participants whether the radiograph was normal or abnormal. Furthermore, if the radiograph was abnormal, the diagnosis was given. Participants had a maximum of 10 seconds to read the feedback page. After 10 seconds, the feedback page closed automatically, and the next radiograph was loaded. In the instruction phase, participants individually watched the video lecture with 18 example chest radiographs. When participants had completed both the practice and the instruction phases, they had a short break of five minutes.

After the break, participants entered the test phase, in which participants had a maximum of 90 seconds to interpret and report every radiograph. After 90 seconds, the next case was automatically loaded. Participants did not receive any feedback about interpreted images during the test phase.

ANALYSES

For the analyses, 2x2 analyses of variance (ANOVAs) were performed with the factors instructional sequence (to practice-first order (inductive) and instruction-first order (deductive)) and proportion (a proportion of 30% normal images versus a proportion of 70% normal images) on the outcome measures of the test phase and practice phase. The sensitivity scores of the test phase and practice phase were negatively skewed; the lowest $Z_{skewness}$-score for the test phase sensitivity was found in the instruction-first (deductive), condition 4, with 70% normal images and was -4.37. The lowest $Z_{skewness}$-score for the practice phase sensitivity was found in the instruction-first order (deductive), condition 3, with 30% normal images and was -2.40. As there is currently no reasonable non-parametric alternative for a 2x2 ANOVA and that ANOVA-analyses are generally robust for skewness, these skewness levels were tolerated. As a measure of effect size, η_p^2 was used, with 0.01 indicating a small effect, 0.06 indicating a moderate effect, and 0.14 indicating a large effect (20, 21).

To analyze differences between the four conditions in case durations for cases divided into correctly identified versus incorrectly identified cases,

a full-factorial binary logistic regression analysis of the practice phase was performed with discrimination score (correct versus incorrect) as the dependent variable and instructional sequence, the proportion, and case duration as independent variables.

RESULTS

Results of the test phase

The descriptors and the results of the 2x2 ANOVA per test-phase measure can be found in Table 2. Furthermore, the descriptors of the test-phase are visualized as violin plots in Figure 3.

Table 2. Test phase measures for each separate condition.

	Practice-first order (inductive)		Instruction-first order (deductive)	
	30% normal (*n*=23)	70% normal (*n*=20)	30% normal (*n*=30)	70% normal (*n*=30)
Variable	*M* ± *SD*	*M* ± *SD*	*M* ± *SD*	*M* ± *SD*
Sensitivity (%)	97.5 ± 5.13	89.1 ± 9.46	96.7 ± 6.51	89.6 ± 8.73
Specificity (%)	52.5 ± 15.3	65.2 ± 14.8	58.3 ± 18.0	72.8 ± 13.3
Diagnostic performance (%)	53.8 ± 18.1	48.9 ± 15.0	52.9 ± 17.9	52.5 ± 13.2
Case duration (s)	35.1 ± 9.31	34.0 ± 10.7	30.4 ± 7.17	29.8 ± 8.25

	2x2 ANOVA		
Variable	Main effect of proportion of normal images	Main effect of instructional sequence	Interaction effect
Sensitivity (%)	$F(1, 99) = 24.97$, $p < .001$, $\eta_p^2 = .20$	$F(1, 99) = 0.02$, $p = .90$, $\eta_p^2 < .001$	$F(1, 99) = 0.17$, $p = .68$, $\eta_p^2 < .001$
Specificity (%)	$F(1, 99) = 20.70$, $p < .001$, $\eta_p^2 = .17$	$F(1, 99) = 5.03$, $p = .03$, $\eta_p^2 = .05$	$F(1, 99) = 0.08$, $p = .77$, $\eta_p^2 < .001$

Table 2 Continues.

Variable	2x2 ANOVA		
	Main effect of proportion of normal images	Main effect of instructional sequence	Interaction effect
Diagnostic performance (%)	$F(1, 99) = 0.67$, $p = .42$, $\eta_p^2 < .001$	$F(1, 99) = 0.18$, $p = .67$, $\eta_p^2 < .001$	$F(1, 99) = 0.47$, $p = .49$, $\eta_p^2 < .001$
Case duration (s)	$F(1, 99) = 1.57$, $p = .21$, $\eta_p^2 < .001$	$F(1, 99) = 9.61$, $p = .003$, $\eta_p^2 = .09$	$F(1, 99) < 0.01$, $p = .95$, $\eta_p^2 < .001$

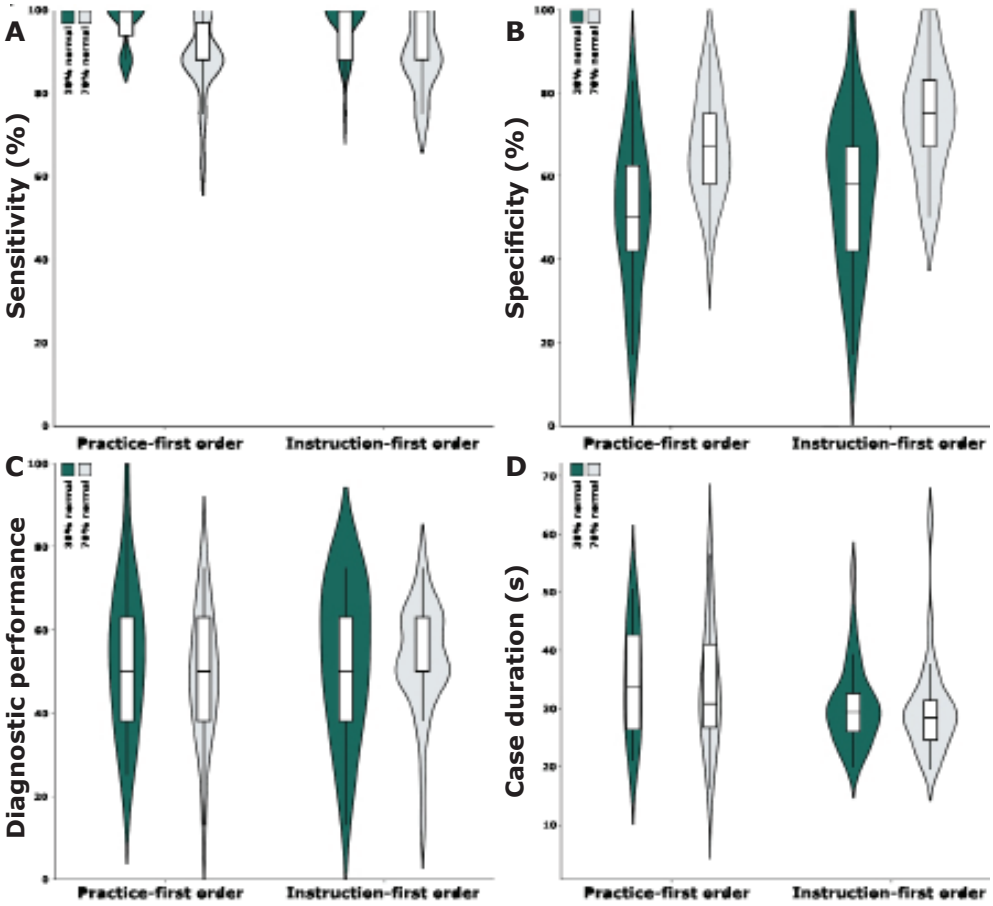
On sensitivity, a main effect of the proportion of normal images was found, in favor of practicing with 30% normal images; the found main effect is in line with hypothesis 1. There was no main effect of sequence. No significant interaction effect between proportion and instructional sequence was found.

On specificity, a main effect of the proportion of images was found in favor of practicing with 70% normal images; the found main effect is in line with hypothesis 1. Furthermore, a main effect of sequence, now in favor of the instruction-first (deductive), conditions 3 and 4, was found. No significant interaction between the proportion of normal images and sequence was found.

On diagnostic performance, no main effect of the proportion was found, which contrasted with hypothesis 1. There was no main effect of sequence, which contrasted with hypothesis 2. No significant interaction effect between proportion and sequence was found.

On average case duration, no main effect of proportion was found. A significant main effect of instructional sequence was found; the average case duration was higher in the practice-first (inductive), conditions 1 and 2. No significant interaction effect between proportion and sequence was found.

*Figure 3. Violin plots of the outcome measures of the test phase per condition. **



* Violin plots represent a regular box plot with 95% confidence intervals, median and interquartile range surrounded by a rotated kernel density plot. A: Sensitivity, B: Specificity, C: Diagnostic performance (%), and D: Average case duration.

The time limit for interpreting cases was reached in five out of 400 cases for the practice-first (inductive), condition 1 with 30% normal images, five out of 460 cases for the group practice-first (inductive), condition 2 with 70% normal images; eight out of 600 cases for the group instruction-first (deductive), condition 3 with 30% normal images, and seven out of 600 cases for the instruction-first (deductive), condition 4 with 70% normal images. The number of cases in which the time limit was reached did not differ between the four conditions, X^2 (3, n = 2060) = .15, p = .99.

Results of the practice phase

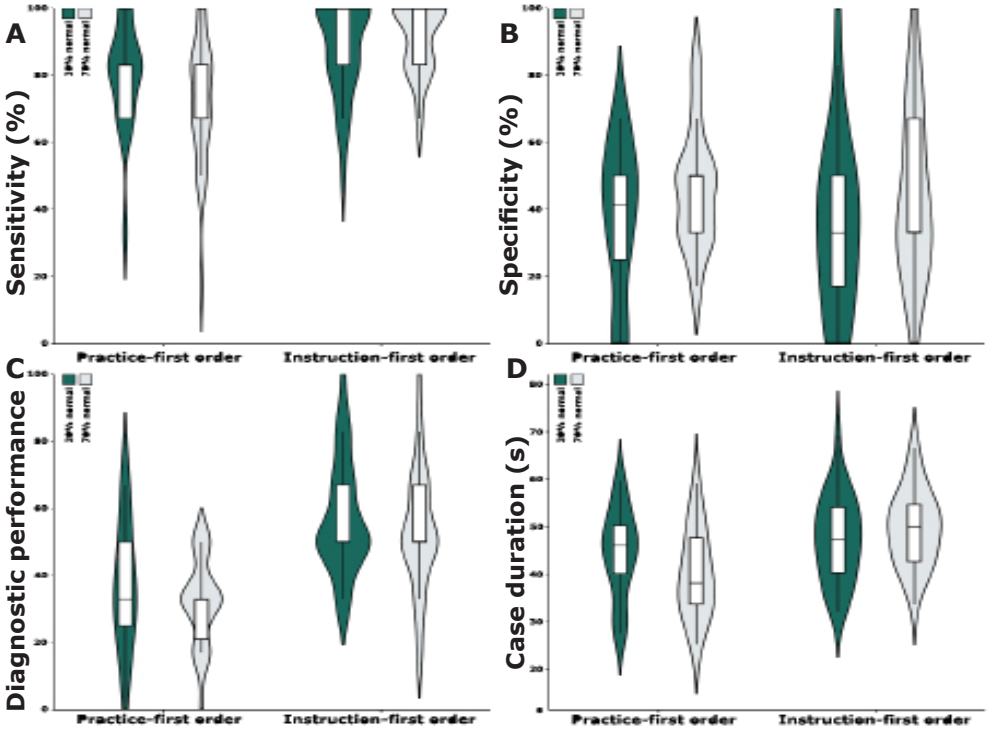
The descriptors and the results of the 2x2 ANOVA of the practice phase can be found in Table 3. Furthermore, the descriptors of the practice phase measures are visualized as violin plots in Figure 4.

Table 3 Descriptives and results of practice-phase measures.

	Practice-first order (inductive)		Instruction-first order (deductive)	
	30% normal (*n*=23)	70% normal (*n*=20)	30% normal (*n*=30)	70% normal (*n*=30)
Variable	*M* ± *SD*	*M* ± *SD*	*M* ± *SD*	*M* ± *SD*
Sensitivity (%)	80.0 ± 15.9	71.7 ± 19.7	88.9 ± 14.0	92.2 ± 10.0
Specificity (%)	36.7 ± 22.7	45.6 ± 17.6	35.0 ± 25.2	44.4 ± 24.5
Diagnostic performance (%)	36.7 ± 18.4	31.9 ± 13.2	57.2 ± 16.8	53.3 ± 18.2
Case duration (s)	44.5 ± 9.40	40.1 ± 9.10	47.4 ± 8.82	49.6 ± 7.90

	2x2 ANOVA		
Variable	Main effect of proportion of normal images	Main effect of instructional sequence	Interaction effect
Sensitivity (%)	$F(1, 99) = 0.67$, $p = .41$, $\eta_p^2 = .01$	$F(1, 99) = 23.9$, $p = <.001$, $\eta_p^2 = .20$	$F(1, 99) = 3.73$, $p = .06$, $\eta_p^2 = .04$
Specificity (%)	$F(1, 99) = 4.00$, $p = .048$, $\eta_p^2 = .039$	$F(1, 99) = 0.01$, $p = .76$, $\eta p2 < .001$	$F(1, 99) = 0.002$, $p = .96$, $\eta_p^2 < .001$
Diagnostic performance (%)	$F(1, 99) = 1.65$, $p = .20$, $\eta_p^2 = .02$	$F(1, 99) = 38.8$, $p < .001$, $\eta p2 = .28$	$F(1, 99) = 0.02$, $p = .90$, $\eta_p^2 < .001$
Case duration (s)	$F(1, 99) = 0.35$, $p = .56$, $\eta_p^2 < .001$	$F(1, 99) = 12.6$, $p = .001$, $\eta_p^2 = .11$	$F(1, 99) = 3.59$, $p = .06$, $\eta_p^2 = .04$

*Figure 4. Violin plots of the outcome measures of the practice phase per condition. **



* Violin plots represent a regular box plot with 95% confidence intervals, median and interquartile range surrounded by a rotated kernel density plot. A: Sensitivity, B: Specificity, C: Diagnostic performance (%) and D: Average case duration.

On sensitivity, no main effect of proportion was found. Furthermore, a main effect of instructional sequence, in favor of the instruction first-order, was found. This finding is in line with hypothesis 3. Finally, a marginally significant interaction between proportion and instructional sequence was found, in favor of the group instruction-first (deductive), condition 3, with 30% normal images.

On specificity, a significant main effect of proportion was found, in favor of practicing with 70% normal images. By contrast with hypothesis 3, there was no main effect of instructional sequence. No interaction effect of proportion of normal images and instructional sequence was found.

On diagnostic performance, no main effect of proportion was found. A main effect of instructional sequence in favor of instruction-first (deductive), conditions 3 and 4, was found, in line with hypothesis 3. No significant interaction effect of proportion and instructional sequence was found.

On case duration, no main effect of proportion was found. Unexpectedly and by contrast with hypothesis 4, the participants in the practice-first (inductive) conditions 1 and 2 took less time to complete the practice cases than the participants of the instruction-first (deductive) conditions 3 and 4. A main effect of instructional sequence was found; the average case duration in the instruction-first (deductive), conditions 3 and 4 groups, was higher. Finally, a marginally significant interaction effect was found; the average case duration of the practice-first (inductive), condition 2 group with 70% normal images, was the lowest.

The number of cases in which the time limit was reached per condition can be found in table 4. The time limit for interpreting cases was more often reached in the instruction-first (deductive), conditions 3 and 4 groups, X^2 (1, n = 2060) = 8.3, p = .004. The number of cases in which the time limit for reading the feedback was reached did not differ between the instructional sequence, X^2 (1, n = 2060) = 1.1, p = .30.

Table 4. Occurrence of time limits during the practice phase per condition.

	Practice-first order (inductive)		Instruction-first order (deductive)	
Practice phase time limits	30% normal (*n*=460)	70% normal (*n*=400)	30% normal (*n*=600)	70% normal (*n*=600)
Time limit interpreta-tion	34	27	74	56
Time limit feedback	2	3	2	10

Occurrence of productive failure during the practice phase

The average case durations for correct and incorrect interpretations of the 12 identical cases during the practice phase and the results of the binary logistic regression can be found in Table 5.

The binary logistic regression analysis showed that in both instruction-first (deductive), conditions 3 and 4, participants took longer to identify cases than in both practice-first (inductive), conditions 1 and 2. Furthermore, a main effect of case duration was found, indicating that correctly identified cases were interpreted faster than incorrectly identified cases.

By contrast with hypothesis 5, all two-way and three-way interaction terms were non-significant, indicating that the found main effects of instructional sequence and case duration were similar for all conditions.

Table 5. Results of the binary logistic regression with correctly identified cases as outcome variable.

	Correctly identified			
	No		Yes	
	n	*M* ± *SD*	*n*	*M* ± *SD*
Practice first, 30% normal	114	44.0 ± 22.9	139	38.0 ± 17.9
Practice first, 70% normal	100	51.1 ± 23.5	120	40.2 ± 17.8
Instruction first, 30% normal	114	62.0 ± 18.9	216	44.9 ± 19.2
Instruction first, 70% normal	137	58.0 ± 18.6	193	42.2 ± 19.3

Table 5 Continues.

Binary logistic regression analysis	*B* (*SE*)	*df*	*p*	*OR* *(95% CI)*
Intercept	1.32 (.34)	1	<.001	3.73
Instructional sequence	1.08 (.48)	1	.025	2.93 (1.14-7.51)
Prevalence	-0.53 (.40)	1	.24	0.59 (0.25-1.41)
Case duration	-0.025 (.0070)	1	<.001	0.98 (0.96-0.99)
Instructional sequence * Prevalence	1.16 (.68)	1	.089	3.19 (0.84-12.2)
Instructional sequence * Case duration	-0.016 (.0090)	1	.081	0.98 (0.97-1.00)
Prevalence * Case duration	0.01 (.0090)	1	.26	1.01 (0.99-1.03)
Instructional sequence * Prevalence * Case duration	-0.014 (.013)	1	.28	0.99 (0.96-1.01)

**SE*: standard error, *OR*: odds ratio, *CI*: Confidence Interval
** Note: χ^2 (7) = 132.66, R^2 = .11 (Cox & Snell), .15 (Nagelkerke)

DISCUSSION

In this experiment, the proportion of normal images during a practice phase and the instructional sequence of medical image perception training were manipulated. The effect of changing the proportion of normal images, previously found by Pusic et al. (2) in a sample of residents, was replicated in our sample of medical students. In line with hypothesis 1, sensitivity scores were highest in the conditions with a low proportion of normal images, and specificity scores were highest in the conditions with a high proportion of normal images. It was thus found that students who train with more normal images are less likely to make false-positive errors (reporting abnormalities that are not present), whereas students training with mostly abnormal images are less likely to miss abnormalities, a phenomenon known as a 'criterion shift' (22, 23). One of the first and most important steps in interpreting medical images is the categorization of the image into

normal or abnormal (24, 25). For this categorization, a decision criterion is used, which is influenced by previous experiences (24). Medical image perception training is generally the first experience that students have of interpreting medical images. A mismatch between the prevalence of abnormalities in training (2) and medical images in everyday clinical practice (7-9) can easily result in students being trained with a suboptimal criterion. Our study shows that a short 20-item training session can already have an impact on this criterion (26).

With regard to the effects of instructional sequences on performance measures, the deductive sequence conditions (3 and 4) led to higher student performance scores than the inductive sequence conditions (1 and 2). The participants in the deductive conditions (3 and 4) scored higher on specificity than the participants in the inductive conditions (1 and 2). This finding contrasts with hypothesis 2. In addition, participants in deductive conditions (3 and 4) had a significantly lower average case duration during the test. Therefore, the participants in the deductive conditions (3 and 4) were not only better in correctly identifying the normal images, but were also faster in their interpretation.

By contrast with hypothesis 1, no effect of instructional sequence was found on sensitivity. This analysis may have been influenced by the high test-phase sensitivity scores. As sensitivity was high in all four conditions, a ceiling effect might have occurred. The sensitivity scores of the practice phase were lower in the inductive and deductive conditions than the sensitivity scores of the test phase. In the practice phase, indeed, a significant effect in favor of the deductive conditions (3 and 4) was found.

A closer look at the results of the practice phase can provide more insights into the effects of instructional sequence on students` learning. In line with hypothesis 4, the participants in the inductive conditions (1 and 2) scored lower on sensitivity, specificity, and diagnostic performance. The students in the inductive conditions were supposed to use the practice cases to develop their own solutions and were thus expected to make more mistakes during the practice phase. However, by contrast with hypothesis 4, the students in the inductive conditions (1 and 2) took less time to complete the practice cases. This finding suggests that they did not explore the cases in enough depth. The inductive approach might, therefore, not have led to productive failure during the practice phase but to *unproductive* failure.

Furthermore, the binary logistic regression analysis revealed that students in all four conditions needed more time for cases they incorrectly interpreted compared with cases they correctly interpreted. This finding indicates that productive failure probably occurred in all four conditions and not only in the inductive conditions (1 and 2). Invoking productive failure may thus not be confined to inductive approaches, and research on other incentives to invoke productive failure is therefore advised.

The lack of increased productive failure in the inductive conditions (1 and 2) is also reflected by the diagnostic performance scores of the test phases. No effect of sequence was found, by contrast with hypothesis 2. One of the claims for the use of inductive approaches is that they lead to a deeper understanding of the problem. In this study, no evidence for this claim was found. A deductive approach is advocated for learners who already have some experience in the task (27). These third-year medical students can be considered novices in the task of image interpretation. However, they may already have acquired some knowledge on chest (patho)physiology during their prior medical training. This knowledge basis might possibly have been solid enough for students to benefit from the deductive approach. Inductive approaches are traditionally advised for the educational experiences of learners confronted with a completely novel task (27). Less experienced students, such as first-year medical students, might have profited from an inductive approach, and replication of this research with less experienced students is therefore advised.

A theoretical pitfall of a criterion shift used should be considered. Because of current educational practice, students are more likely to make false-positive interpretations. False-positive and false-negative interpretations of medical images have different consequences for patient outcomes. False-positive interpretations may lead to unnecessary diagnostic procedures, whereas false-negative interpretations may lead to potentially life-threatening delays in diagnoses (26). However, novices generally make more false-positive errors than false-negative errors. This is even the case for the interpretation of images with a much lower prevalence of diseases than chest X-rays, such as the prevalence of breast abnormalities in breast cancer screening programs (28, 29). It is unlikely that a shift in the criteria used by novices would lead to an increase in false-negative interpretations. It is therefore advised to take the prevalence of diseases into account when developing training.

With the limited time that faculty members have available for medical image perception training (6), the question arises: How should students be trained to identify diverse pathologies while still developing realistic criteria (28)? Additional e-learning modules containing large image banks with the proportion of abnormalities seen in everyday clinical practice are advised.

Additionally, the use of a deductive approach is advised. In many faculties, medical image perception training is provided when students have already acquired some knowledge of anatomy and (patho)physiology (30).

Some limitations of this research are worth considering. In this research, learning outcomes were directly measured, and no measures of retention of knowledge were used. Inductive sequences are also advocated to enhance retention of knowledge, yet evidence for this claim is still limited (12). Further research to elucidate the effects of early practice is therefore recommended. Furthermore, participants were asked to make a clear distinction between normal and abnormal, whereas everyday medical practice is not that black and white. In everyday medical practice, abnormal images still predominantly consist of normal areas, and normal images can contain aberrations, which could be abnormal in some clinical cases. To ensure a clear cut-off between normal and abnormal in this study, only images with apparent abnormalities were used, and clinical information was not provided to participants.

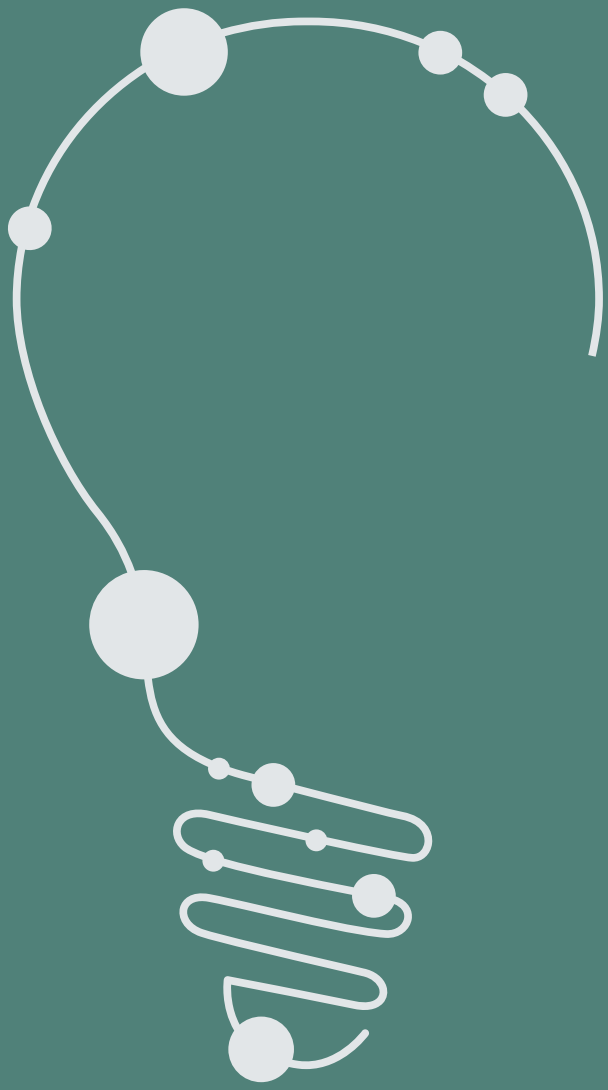
CONCLUSION

On immediate post-testing, a deductive approach for training third-year medical students to interpret radiographs yielded better results than an inductive approach in discerning normal from abnormal images. Furthermore, it was shown that the proportion of normal images in a training situation impacts the criteria students use to categorize normal and abnormal medical images. In many medical situations, the prevalence of diseases is low, and the sensitivity and specificity trade-off should be an important consideration in training design.

REFERENCES

1. Iglehart JK. The new era of medical imaging--progress and pitfalls. The New England journal of medicine. 2006;354(26):2822-8.
2. Pusic MV, Andrews JS, Kessler DO, Teng DC, Pecaric MR, Ruzal-Shapiro C, et al. Prevalence of abnormal cases in an image bank affects the learning of radiograph interpretation. Med Educ. 2012;46(3):289-98.
3. Gegenfurtner A, Kok E, van Geel K, de Bruin A, Jarodzka H, Szulewski A, et al. The challenges of studying visual expertise in medical image diagnosis. Med Educ. 2017;51(1):97-104.
4. Kok EM, van Geel K, van Merriënboer JJG, Robben SGF. What We Do and Do Not Know about Teaching Medical Image Interpretation. Frontiers in Psychology. 2017;8(309).
5. van der Gijp A, Ravesloot CJ, Jarodzka H, van der Schaaf MF, van der Schaaf IC, van Schaik JPJ, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education. 2017;22(3):765-87.
6. Zafar AM. Radiology: an underutilized resource for undergraduate curricula. Med Teach. 2009;31(3):266.
7. Verma V, Vasudevan V, Jinnur P, Nallagatla S, Majumdar A, Arjomand F, et al. The utility of routine admission chest X-ray films on patient care. Eur J Intern Med. 2011;22(3):286-8.
8. Marcolino MS, Palhares DM, Alkmim MB, Ribeiro AL. Prevalence of normal electrocardiograms in primary care patients. Rev Assoc Med Bras (1992). 2014;60(3):236-41.
9. Ng JJ, Taylor DM. Routine chest radiography in uncomplicated suspected acute coronary syndrome rarely yields significant pathology. Emergency medicine journal : EMJ. 2008;25(12):807-10.
10. Mylopoulos M, Brydges R, Woods NN, Manzone J, Schwartz DL. Preparation for future learning: a missing competency in health professions education? Med Educ. 2016;50(1):115-23.
11. van Merriënboer JJG, Sweller J. Cognitive load theory and complex learning: Recent developments and future directions. Educational Psychology Review. 2005;17(2):147-77.
12. Lee HS, Anderson JR. Student learning: what has instruction got to do with it? Annu Rev Psychol. 2013;64:445-69.
13. Kapur M. Productive Failure. Cognition and Instruction. 2008;26(3):379-425.
14. Soderstrom NC, Bjork RA. Learning Versus Performance:An Integrative Review. Perspectives on Psychological Science. 2015;10(2):176-99.
15. Kapur M. Productive Failure in Learning Math. Cognitive Science. 2014;38(5):1008-22.
16. Kok E, Jarodzka H, Bruin AB, Bin Amir H, Robben SGF, Merrienboer JG. Systematic viewing in radiology: seeing more, missing less? Adv Health Sci Educ Theory Pract. 2015.
17. van Geel K, Kok EM, Dijkstra J, Robben SG, van Merrienboer JJ. Teaching Systematic Viewing to Final-Year Medical Students Improves Systematicity but Not Coverage or Detection of Radiologic Abnormalities. Journal of the American College of Radiology : JACR. 2017;14(2):235-41.
18. Revelle W, Zinbarg RE. Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. Psychometrika. 2009;74(1):145-54.
19. Qualtrics. Provo, USA. 2016 [Available from: https://www.qualtrics.com/.
20. Field AP. Discovering Statistics Using SPSS: Sage; 2009.
21. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Taylor & Francis; 2013.
22. Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. Journal of experimental psychology General. 2007;136(4):623-38.
23. Treisman M, Williams TC. A theory of criterion setting with an application to sequential dependencies. Psychological Review. 1984;91(1):68-111.
24. Maddox WT. Toward a unified theory of decision criterion learning in perceptual categorization. Journal of the Experimental Analysis of Behavior. 2002;78(3):567-95.
25. Krupinski EA. Current perspectives in medical image perception. Attention Perception & Psychophysics. 2010;72(5):1205-17.
26. Berlin L. Radiologic errors and malpractice: a blurry distinction. AJR Am J Roentgenol. 2007;189(3):517-22.
27. van Merriënboer JJG, Kirschner PA. Ten Steps

to Complex Learning: A Systematic Approach to Four-component Instructional Design: Routledge; 2018.
28. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. 2009;253(3):632-40.
29. Alberdi RZ, Llanes AB, Ortega RA, Exposito RR, Collado JM, Verdes TQ, et al. Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. European radiology. 2011;21(10):2083-90.
30. Linaker KL. Radiology Undergraduate and Resident Curricula: A Narrative Review of the Literature. Journal of Chiropractic Humanities. 2015;22(1):1-8.



Chapter 3

Teaching systematic viewing to final-year medical students improves systematicity but not coverage or detection of radiologic abnormalities

Koos van Geel, Ellen M. Kok, Joost Dijkstra, Simon G.F. Robben, Jeroen J.G. van Merriënboer

Journal of the American College of Radiology: 2017;14(2):235-41.

ABSTRACT

Purpose

Systematic viewing of images is widely advocated in radiology; it is expected to lead to complete coverage of images and consequently, more detection of abnormalities. Evidence on the efficacy of teaching systematic viewing to students is conflicting. The aim of this study was to investigate the effects of teaching systematic viewing to final-year medical students on systematicity of viewing behavior, coverage of the image, and detection.

Methods

Final-year students (n = 60) viewed 10 chest-radiographs in a first series before training and another 10 radiographs in a second series after training. Between series, students were taught basic chest-radiograph viewing, either in a systematic or a nonsystematic manner. With eye-tracking, systematicity (Levenshtein distances), coverage (percentage of an image viewed), and detection (sensitivity and specificity) were measured.

Results

In a mixed 2x2 design, significantly higher sensitivity was found after training compared with before training (F(1,55) = 6.68, p = .012, η_p^2 = .11) but no significant effect for type of training (F(1,55) = 1.24, p = .30) and no significant interaction effect (F(1,55) = 0.12, p = .73). Thus, training in systematic viewing was not superior to training in nonsystematic viewing. A significant interaction of training type and time of viewing was found on systematicity, (F (1,49) = 20.0, p < .01, η_p^2= .29) in favor of the systematic viewing group. No significant interaction was found for coverage (F(1,49) = 0.43, p = .51) or specificity (F(1,55) = 0.124, p = .73).

Conclusion

Both training types showed similar increases in sensitivity. Therefore, it might be advisable to pay less attention to systematic viewing and more attention to content, such as the radiological appearances of diseases.

INTRODUCTION

A systematic approach is widely recommended to medical students when they are taught to interpret radiological abnormalities (1-3). Such systematic viewing approaches may differ in the order in which anatomical structures should be looked at, but all concur that students need to adhere to one specific order for all images. The principle behind pursuing the same specific order is that students will be less likely to overlook anatomical structures in their viewing process and will, therefore, be most complete. By completely covering images, medical students are expected to miss fewer abnormalities. Although it is common practice in radiology departments to teach novices a systematic approach, little research has been performed on its efficacy.

The effects of systematic viewing on detection were investigated by Peterson (4) and Auffermann et al. (5). Peterson found that students who used a complete but nonsystematic search pattern performed significantly better than students who used any other search pattern. Peterson's study, however, had only an observational design, and thus the effects of systematic viewing training on detection remained unknown. Furthermore, search patterns and completeness were deduced from think-aloud data rather than from more objective data. Using think-aloud data as a measure of viewing behavior carries the assumption that one could objectively report where one is looking, which is an assumption that does not hold (6). To objectively measure viewing behavior, the movements of the eyes need to be captured, which can be done by measuring participants` eye movements with eye-tracking apparatus (7).

Auffermann et al. (9) investigated the effect of training in systematic viewing on physician assistant trainees evaluating chest radiographs. They found that trainees who participated in the training detected significantly more abnormalities in comparison with the control group. Unfortunately, the control group of this study did not have equal exposure to training in chest radiographic interpretation. Thus, it is unclear whether the increase in detection was the result of the greater educational exposure (3, 8) or the result of the instruction to evaluate images systematically. Furthermore, Auffermann et al. (5) did not use measures for search patterns or coverage in their methodology, and the effects of training on search patterns are unknown.

Thus, to establish the effectiveness of training in systematic viewing, research is required that uses objective (eye-tracking) data to quantify systematic viewing. Furthermore, the effectiveness of training in systematic viewing needs to be established against training in nonsystematic viewing that has equal educational exposure. In this study, we compared a group of final-year medical students who received training in systematic viewing with a group that received similar training that did not focus on systematic viewing. Eye movements were measured using eye-tracking methodology. The aim of this study was to answer the following research questions:

1. Does the detection of abnormalities increase after training in systematic viewing when final-year medical students view chest x-rays?
2. Do eye movements change after training in systematic viewing, showing increased systematicity and coverage when final year medical students view chest x-rays?

METHODS

Participants

Final-year medical students (n = 60, 73% female; mean age, 24.8 ± 1.54 years) participated in this experiment. All students were recruited from Maastricht University Medical Center or affiliated hospitals. Students were recruited via the electronic learning environment of Maastricht University

All participants had some experience in viewing chest radiographs during their prior clinical rotations but had not received any formal training. Students who had followed an elective chest radiology rotation or who were performing final-year internships in a radiology department were not included. Participants were randomly assigned to one of the two groups; 31 were allotted to the systematic viewing group and 29 to the nonsystematic viewing group. The participants received a €10 gift voucher as a reward.

MATERIALS

Apparatus

Eye movements were measured using an SMI RED remote eye tracker (SensoMotoric Instruments, Teltow, Germany). The head movements of participants were not physically restricted. However, to ensure optimal data quality, participants were instructed to avoid head movements as much as possible. The sampling rate was set to 250 Hz, and the eye movements

of participants` right eye were used. The images were shown on a Dell 22-inch liquid crystal display, using a resolution of 1,650 x 1,080 pixels. Before the start of the first (pretraining) and the second (posttraining) series of images, the eye tracker was calibrated using a nine-point calibration. Calibration was repeated until a deviation of less than 1º of visual angle on both the x-axis and y-axis was acquired. Eye-tracking data from nine participants were excluded from the analysis due to insufficient data quality (i.e., the threshold of 1º of visual angle could not be reached). Data were analyzed using IBM SPSS Statistics 21 (IBM, Amsterdam, the Netherlands).

Radiologic images

In this study, chest radiographs were used. Chest radiographs not only account for considerate amounts of work in every radiology department (9), but viewing them is also difficult to master (10). Therefore, using chest radiographs would minimize potential ceiling effects. To ensure the inclusion of images with distinct pathology and distinct normal images, all chest radiographs were individually evaluated by two senior radiologists. Images were included only when the radiologists agreed in their evaluations. All images were stripped of any identifying information. Of the total set of 20 chest radiographs, 17 contained two or more abnormalities, and the other three were normal. The number of abnormalities was 56 in total: 33 in the pretraining image series and 23 in the posttraining image series (see Figure 1). The abnormalities on the images differed in shape, size, and location and were manifestations of the following diseases: pneumonia, atelectasis, cardiomegaly, pleural effusion, lung tumor, pneumothorax, lung emphysema, and hilar lymphadenopathy.

Training in Systematic and Nonsystematic viewing

Two instructional videos were used in our experiment to teach participants to use either a systematic viewing approach or a nonsystematic viewing approach. The videos were previously used in an experiment by Kok et al. (13). The training videos differed only with respect to the advocated viewing approach. Participants of both groups hence saw a video of approximately 30 min long in which the basics of chest radiograph interpretation were explained, with the appearances of the previously mentioned cardiopulmonary diseases. Participants saw the video only once and were not allowed to stop, rewind, or fast-forward the video. Moreover, they were not allowed to make notes. In the video for the systematic viewing group, a systematic viewing approach was encouraged. In contrast, the participants of the

nonsystematic viewing group were discouraged from using a fixed order during their viewing process. Therefore, they were instructed to view only whatever primarily drew their attention.

MEASURES

Detection measures

Two measures of detection were used, sensitivity and specificity. To calculate sensitivity and specificity, participants were asked to click on all abnormalities they saw. Sensitivity was defined as the number of correctly clicked abnormalities divided by the total number of abnormalities of an image series. Sensitivity was calculated per image and then averaged over all images of a series. Specificity was defined as the number of images of an image series where the participant did not click on any healthy tissue divided by the total number of images of an image series. The detection measures of three participants were not registered due to technical difficulties and were excluded from the analysis of sensitivity and specificity.

Systematicity and coverage measures

The minimal fixation duration was set to 100 ms. To measure systematicity, Levenshtein distances were calculated (11), which is the most used measure in eye-tracking research for comparing the similarity of the eye movements on two images (7). Specifically, eye movements on one image can be understood as a string or chain of fixations. By comparing the chain of fixations on one image with the chain of fixations on another image, the similarity of viewing processes per participant can be calculated. To construct such a chain of fixations, we superimposed a 7x7 grid on each image. We then determined which cells were fixated and in which order. All grid cells were subsequently ranked, based on the time to the first fixation. Next, the minimal number of modifications (deletions, insertions, or substitutions) were calculated that were required for the chain of grid fixations of the second image to become the first. Finally, this number was divided by the maximum number of fixated grids, which resulted in the Levenshtein distance. The Levenshtein distances between each pair of images in each series were computed and averaged per participant per image series. Fewer modifications result in a lower Levenshtein distance and indicate higher systematicity.

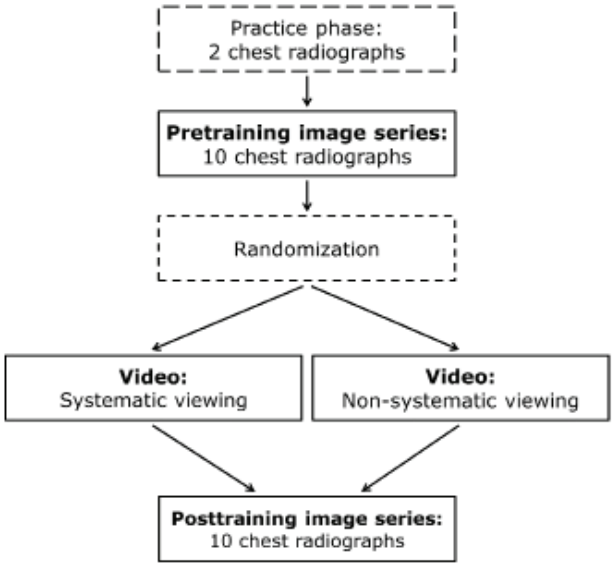
To measure coverage, a 7x7 grid was again superimposed on each image. Coverage was subsequently calculated as the number of grid cells a

participant had fixated on at least once, divided by the total number of cells. Coverage was calculated per image and averaged for each of the image series.

Procedure

The procedure of this experiment is delineated in Figure 1. Prior to the experiment, participants provided written informed consent. Subsequently, the instructions for the pretraining image series were practiced during a short practice phase with two chest radiographs. For the pretraining image series, participants were asked to view the radiographs as they would do during their clinical rotations. When they finished viewing a radiograph, participants pressed the space bar, after which a new screen was opened, in which they could type their findings. Furthermore, the first image of this series was used to validate the eye-tracker calibration. The sensitivity, specificity, and eye movements of the first image series were recorded.

Figure 1. Flowchart of the experiment.



After completion of the pretraining image series, participants saw their respective training videos. During the videos, the eye movements of the participants were not recorded. After their training, participants viewed the radiographs of the posttraining image series, and the sensitivity, specificity, and eye movements were again recorded. Contrary to the pretraining image series, they were not asked to view as they would do during their clinical

rotations, but to follow the instructions of their respective systematic or nonsystematic training. There were no time restrictions. Each participant performed the experiment in an individual session. Only after completion of the whole experiment could the diagnoses of the images be provided at the request of the participant.

ANALYSES

To analyze the data, 2x2 analyses of variance with the factors type of training (systematic viewing versus nonsystematic viewing) and time of viewing images (posttraining versus pretraining) were used to identify intergroup differences on systematicity, coverage, and the detection measures. Each analysis of variance tested three effects: two main effects of each separate factor and one interaction effect between the two factors. The main effect of time of viewing is an estimation of the change from pretraining to posttraining. The main effect of type of training is an estimation of the overall difference between the nonsystematic and the systematic training group. The interaction effect is an estimation of the relation between the change over time and the difference between the two groups: for example, does the change over time of sensitivity significantly differ between the systematic and the nonsystematic viewing group? Therefore, the research questions refer to the interaction effect. The η_p^2 statistic was calculated to measure effect size, with .01 indicating a small effect, .06 indicating a moderate effect, and .14 indicating a large effect (12).

ETHICAL REVIEW

Ethical approval was received from the ethical review board of the Dutch Association for Medical Education (NVMO-ERB), file number 334.

RESULTS

Does the detection of abnormalities increase after training in systematic viewing when final-year medical students view chest x-rays?

The detection measures of the pretraining and posttraining image series can be found in Table 1 and are further visualized in Figure 2. No significant interaction effect between the type of training and time viewing of the posttraining images series compared with the pretraining image series was found on sensitivity ($F(1,55) = 0.12$, $p = .73$, $\eta_p^2 = .002$). However, there was a main effect of time of viewing, showing significantly higher sensitivity in the posttraining image series compared to the pretraining image series for both groups ($F(1,55) = 6.68$, $p = .012$, $\eta_p^2 = .11$). Furthermore, there

was no significant main effect of training, ($F(1,55) = 1.24$, $p = .30$, $\eta_p^2 = .022$). Thus, training in systematic viewing did not yield more improvement in sensitivity than training in nonsystematic viewing.

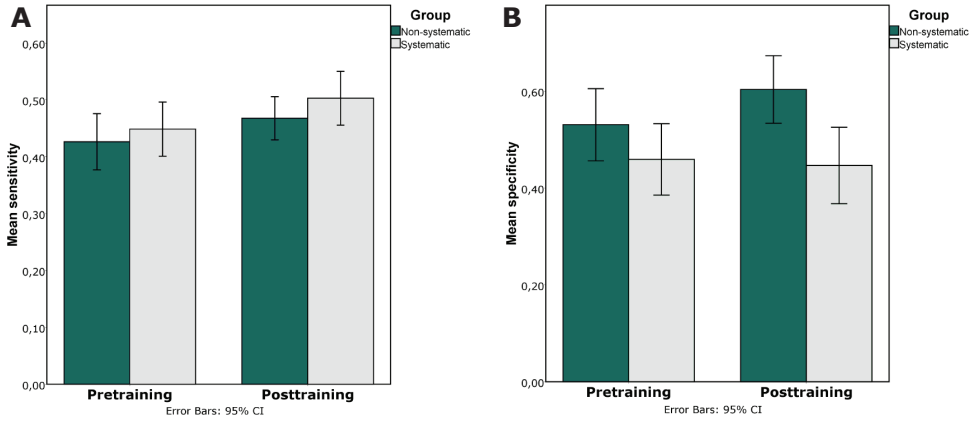
Table 1. Sensitivity and specificity of the pretraining and posttraining image series in the nonsystematic and systematic viewing group.

	Pretraining			
	Nonsystematic		Systematic	
Variable	*N* ± *SD*	*M* (%) ± *SD*	*N* ± *SD*	*M* (%) ± *SD*
Sensitivity*	14.1 ± 4.1	42.6 ± 12.5	14.8 ± 4.2	44.8 ± 12.7
Specificity**	5.3 ± 1.9	53.0 ± 18.9	4.6 ± 2.0	45.9 ± 19.7

	Posttraining			
	Nonsystematic		Systematic	
Variable	*N* ± *SD*	*M* (%) ± *SD*	*N* ± *SD*	*M* (%) ± *SD*
Sensitivity*	10.8 ± 2.2	46.8 ± 9.61	11.6 ± 2.9	50.3 ± 12.7
Specificity**	6.0 ± 1.8	60.3 ± 17.6	4.5 ± 2.1	44.7 ± 21.1

* Numbers (*N* ± *SD*) of sensitivity represent average found abnormalities per image series for both groups
** Numbers (*N* ± *SD*) of specificity represent average correctly identified images per image series for both groups

Figure 2. Pretraining and posttraining sensitivity (A), and specificity (B) of the systematic and nonsystematic group.



No significant interaction between type of training and time of viewing was found on specificity ($F(1,55) = 0.124$, $p = .73$, $\eta_p^2 = .002$). Thus, training in systematic viewing did not yield more improvement in specificity than training in nonsystematic viewing. There was no main effect of time of viewing on specificity ($F(1,55) = 0.847$, $p = .36$, $\eta_p^2 = .015$). There was, however, a significant effect of type of training ($F(1,55) = 8.23$, $p < .01$, $\eta_p^2 = .13$): participants in the nonsystematic group had a higher overall specificity compared to the systematic group.

Do eye movements change after training in systematic viewing, showing increased coverage and systematicity when final-year medical students view chest x-rays?

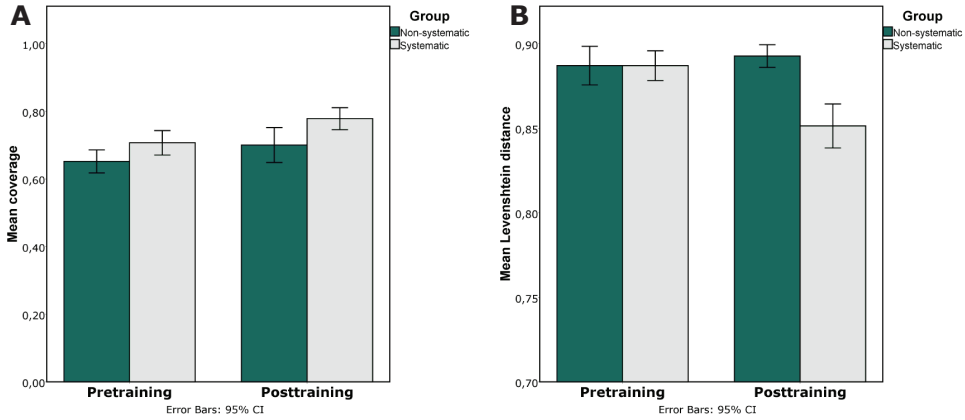
The coverage and systematicity of both groups of the pretraining image series and the posttraining image series can be found in Table 2 and are further visualized in Figure 3.

The interaction effect of type of training with time of viewing on coverage was not significant ($F(1,49) = 0.43$, $p = .51$, $\eta_p^2 = .01$). Thus, the increase in coverage of the systematic group did not significantly differ from the increase in the nonsystematic group. There was a main effect of time of viewing on coverage ($F(1,49) = 14.8$, $p < .01$, $\eta_p^2 = .23$), indicating that coverage increased after the training in both groups. There was also a main effect of type of training on coverage ($F(1,49) = 6.80$, $p = .012$, $\eta_p^2 = .12$), showing higher coverage for the systematic viewing group than the nonsystematic viewing group.

Table 2. Coverage and Levenshtein distances of the nonsystematic and systematic viewing group during the pretraining and posttraining image series.

	Pretraining		Posttraining	
	Non-systematic	Systematic	Non-systematic	Systematic
Variable	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Coverage	65.7% (8.00)	70.6% (10.0)	71.6% (12.1)	77.8% (9.16)
Levenshtein distance	.89 (.03)	.89 (.022)	.89 (.015)	.85 (.032)

Figure 3. Pretraining and posttraining coverage (A) and Levenshtein Distance (B) of the nonsystematic and systematic group.



Furthermore, there was a significant interaction between type of training and time of viewing on systematicity, ($F(1,49) = 20.0$, $p < .01$, $\eta_p^2 = .29$), indicating that the systematic group became significantly more systematic compared to the nonsystematic group after the training in systematic viewing. Furthermore, there was a main effect of time of viewing on systematicity, ($F(1,49) = 11.3$, $p < .01$, $\eta_p^2 = .19$), and there was a significant effect of training on systematicity: participants in the systematic group were significantly more systematic than participants in the nonsystematic group ($F(1,49) = 12.5$, $p < .01$, $\eta_p^2 = .20$).

DISCUSSION

We investigated the effects of training in systematic viewing on detection, systematicity, and coverage of radiological abnormalities among final-year medical students. With regard to the first research question, whether detection increases after training in systematic viewing, sensitivity in both groups increased significantly, and to the same degree, after their respective training. The types of training in this research differed only with respect to how the radiological content was taught; training in systematic viewing did not prove to be superior to training in nonsystematic viewing. Therefore, the findings of this research indicate that when teaching radiology to medical students, training should primarily emphasize on radiological content, such as radiological manifestations of diseases, rather than on training search patterns.

With regard to the second research question, whether eye movements change after training in systematic viewing, showing increased systematicity and coverage, a significant interaction effect on systematicity was found, indicating that the systematic viewing group significantly increased their systematicity, whereas the systematicity of the nonsystematic group remained stable. We thus conclude that both groups were able to adapt their viewing behavior on the basis of their respective instructions and thus followed the provided instructions. Moreover, no significant interaction effect on coverage was found, although both groups increased significantly in coverage after their training. This indicates that to increase coverage, training in systematic viewing is also not favorable to training in nonsystematic viewing, in which students were instructed to look at whatever drew their attention.

Our finding that training in systematic viewing does not increase detection may seem contradictory to Auffermann et al.'s findings (5). In Auffermann et al.'s study, however, participants of the control group had less exposure to chest radiographs, and the systematic viewing group was trained more extensively than the control group. Training and exposure to radiographs is a strong factor in learning radiology (10). Indeed, in a similar setup to ours, but with third-year medical students, no difference in detection was found between training in systematic viewing and training in nonsystematic viewing (13).

Because training in systematic viewing was not found to be superior to increase detection, the role of training in systematic viewing in radiology education should be further examined. In education, however, not only efficacy but also the preferences of students should be considered. Students prefer their radiology education to include systematic viewing approaches as it might give them guidance (14): many students find it initially difficult to start viewing radiographs as they yet do not know where to begin. Indeed, students consider such approaches valuable when applied (5).

Because students prefer the guidance of systematic viewing in their education, further research should focus on ways to optimize such guidance. Instead of using a lecture or instructional video to teach students how to use a systematic approach, methods that provide more support should be considered. Checklists have the potential to be such a supportive method. Checklists are essentially lists of criteria, organized in a systematic fashion

(15, 16), to ensure that all steps in a complex procedure, such as viewing a radiological image, are considered. Because checklists have already proved their worth for learning in other medical specialties, such as surgery (17), further research on the effectiveness of checklists for learning radiology is implicated (16).

Limitations

The present study has some limitations. First, no clinical information on the images was provided during the experiment. Because clinical information potentially influences viewing behavior (18), and we were interested in the role of training in systematic viewing on viewing behavior, this particular factor was controlled for in this experiment. However, to further unravel the effects of systematic viewing on learning to view radiographs, further research should focus on the combined effects of clinical information and training in systematic viewing on detection and viewing behavior.

Second, the training used in this experiment consisted of instructional videos of approximately 30 min long, which may be too limited to train participants extensively. Although short-term effects were found in this research, the effects of training viewing behavior in the long term are not investigated so far. Further research should investigate the long-term effects of training in systematic viewing on detection and viewing behavior.

TAKE-HOME MESSAGES

• Teaching radiology to final-year medical students increases the detection of abnormalities on chest radiographs.

• Systematic viewing was not found to be superior for the detection of abnormalities.

• Radiology education should emphasize the contents of images, such as the radiological appearances of diseases and variants of normal.

REFERENCES

1. Daffner RH. Clinical Radiology, the Essentials. Lippincott: Williams & Wilkins; 2007.
2. Eastman GW, Wald C, Crossin J. Getting started in clinical radiology from image to diagnosis. Stuttgart; New York: Thieme; 2006.
3. Kourdioukova EV, Valcke M, Derese A, Verstraete KL. Analysis of radiology education in undergraduate medical doctors training in Europe. Eur J Radiol. 2011;78(3):309-18.
4. Peterson C. Factors associated with success or failure in radiological interpretation: diagnostic thinking approaches. Medical Education. 1999;33(4):251-9.
5. Auffermann WF, Little BP, Tridandapani S. Teaching search patterns to medical trainees in an educational laboratory to improve perception of pulmonary nodules. JMIOBU. 2015;3(1):011006-.
6. Ericsson KA, Simon HA. Protocol Analysis: Verbal Reports As Data: Mit Press; 1993.
7. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Weijer J. Eye Tracking: A Comprehensive Guide to Methods and Measures: Oxford University Press; 2011.
8. Sendra-Portero F, Torales-Chaparro OE, Ruiz-Gomez MJ. Medical students' skills in image interpretation before and after training: a comparison between 3rd-year and 6th-year students from two different medical curricula. Eur J Radiol. 2012;81(12):3931-5.
9. Levin DC, Rao VM, Parker L, Frangos AJ. Analysis of radiologists' imaging workload trends by place of service. Journal of the American College of Radiology : JACR. 2013;10(10):760-3.
10. Manning DJ, Ethell SC, Donovan T, Crawford T. How do Radiologists do it? The Influence of Experience and Training on Searching for Chest Nodules. Radiography. 2006;12(2):134-42.
11. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. . Soviet Physics Doklady. 1966;10:707-10.
12. Field AP. Discovering Statistics Using SPSS: Sage; 2009.
13. Kok E, Jarodzka H, Bruin AB, Bin Amir H, Robben SGF, Merrienboer JG. Systematic viewing in radiology: seeing more, missing less? Adv Health Sci Educ Theory Pract. 2015.
14. Subramaniam RM, Beckley V, Chan M, Chou T, Scally P. Radiology curriculum topics for medical students: students' perspectives. Academic radiology. 2006;13(7):880-4.
15. Hales BM, Pronovost PJ. The checklist--a tool for error management and performance improvement. Journal of critical care. 2006;21(3):231-5.
16. Marcovici P, Blume-Marcovici A. Intuition versus rational thinking: psychological challenges in radiology and a potential solution. Journal of the American College of Radiology : JACR. 2013;10(1):25-9.
17. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AH, Dellinger EP, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. The New England journal of medicine. 2009;360(5):491-9.
18. Kundel HL, Wright DJ. The influence of prior knowledge on visual search strategies during the viewing of chest radiographs. Radiology. 1969;93(2):315-20.

Chapter 4

Visual search patterns rapidly change during the first months of radiology residency training; a longitudinal, prospective eye-tracking study

Koos van Geel, Ellen M. Kok, Jeroen H.L.M. Donkers, Ulrich C. Lalji, Diederick C. Niehorster, Simon G.F. Robben, Jeroen J.G. van Merriënboer

Submitted for publication.

ABSTRACT

Background

Reading radiologic images starts with a visual search for abnormalities. Visual searches thus play pivotal roles in the reading processes. Visual search patterns are known to differ between residents and radiologists and may be used to monitor residents' development. However, prior studies were generally cross-sectional and cannot reveal how visual search patterns develop longitudinally. The purpose of this study is to prospectively investigate the longitudinal development of visual search patterns and lesion detection of first-year residents in radiology.

Methods

Radiology residents (n=16) read 20 abnormal and normal chest radiographs (CXRs) in 11 experimental sessions during their first year of residency. Reading time per CXR was recorded. Visual search patterns were measured using eye-tracking technology. The number of fixations, the average fixation duration, the proportion of abnormal dwell time (the sum of fixation durations on abnormalities divided by the sum of total fixation durations), and the mean saccade length per CXR were used as visual search measures. The residents clicked on abnormalities they identified, if any. Sensitivity (proportion of abnormalities clicked on divided by total abnormalities per image), and specificity (normal images not clicked on normal tissue) were used as the lesion detection measures. Data were analyzed using multilevel Cox regression analyses for visual search measures and multilevel logistic regression analyses for sensitivity and specificity.

Results

The reading times were halved during the first four months of training. The number of fixations and the average fixation duration decreased longitudinally, whereas the proportion of abnormal dwell time, and the mean saccade length increased. These findings indicate more efficient visual searches, with pronounced changes occurring during the first four months, whereas sensitivity and specificity remained constant. Visual search patterns differed between abnormal and normal CXRs, indicative of adaptation to image characteristics.

Conclusions

Residents develop more efficient visual search patterns, particularly during the first months of residency, and showed adaptation to image characteristics. The found visual search patterns provide more insights into the development of residents. Eye-tracking technology can foster the monitoring of residents' development.

INTRODUCTION

Reading radiologic images starts with a visual search to detect lesions that are subsequently analyzed (1-4). Visual search patterns are thus fundamental in the reading process. Radiologists have the most efficient visual search patterns and are the most accurate in reading radiological images (5, 6). Novices have to develop efficient visual search patterns in only 4-5 years of residency training (7, 8). It is yet unclear how residents develop visual search patterns over time (3, 9).

Radiologists` visual search patterns have been studied in detail with eye-tracking technology (2, 6, 10). Eye-tracking technology quantifies where, when, and for how long a person has looked (11). Eye movements are divided into fixations and saccades (9, 12-14). Fixations are the moments when the eyes stand relatively still and acquire visual information. Saccades are the rapid movements between fixations when the eyes are functionally blind. It is generally found that radiologists, compared to novices, have lower reading times and a lower number of fixations with shorter durations. Furthermore, radiologists fixate longer on abnormal areas and show longer saccade lengths (2, 6, 10, 15). These findings indicate that radiologists can efficiently guide their visual attention to relevant areas and ignore irrelevant areas (15-17). Taking these extensive differences between radiologists and novices into account, visual search patterns may be used to monitor residents` development of visual skills (7).

Unfortunately, previous eye-tracking studies investigating visual search patterns were generally cross-sectional. They can only provide insights between groups and provide little information on how development over time takes place (9). Furthermore, cross-sectional studies might assume a steady, linear development, an assumption that probably does not hold. Indeed, studies on other aspects of residents' development, such as diagnostic accuracy, found non-linear patterns (8, 18): A logarithmic pattern was found on the longitudinal development of residents' diagnostic accuracy (8). Another longitudinal study of first-year residents evaluating ankle radiographs even showed temporary decreases in diagnostic accuracy (18). Such non-linear developmental trajectories make it harder to estimate residents' actual progress and thus call for additional monitoring. To use visual search patterns to monitor resident' development, it is essential to unravel how these patterns change longitudinally.

Our prospective, longitudinal eye-tracking study investigated the development of residents' visual search patterns and lesion detection. First-year residents read 20 chest radiographs (CXRs) in 11 experimental sessions spread over one year. The following research questions are addressed:

1. How do visual search patterns change over time in the first year of residency?
2. How does lesion detection change over time in the first year of residency?

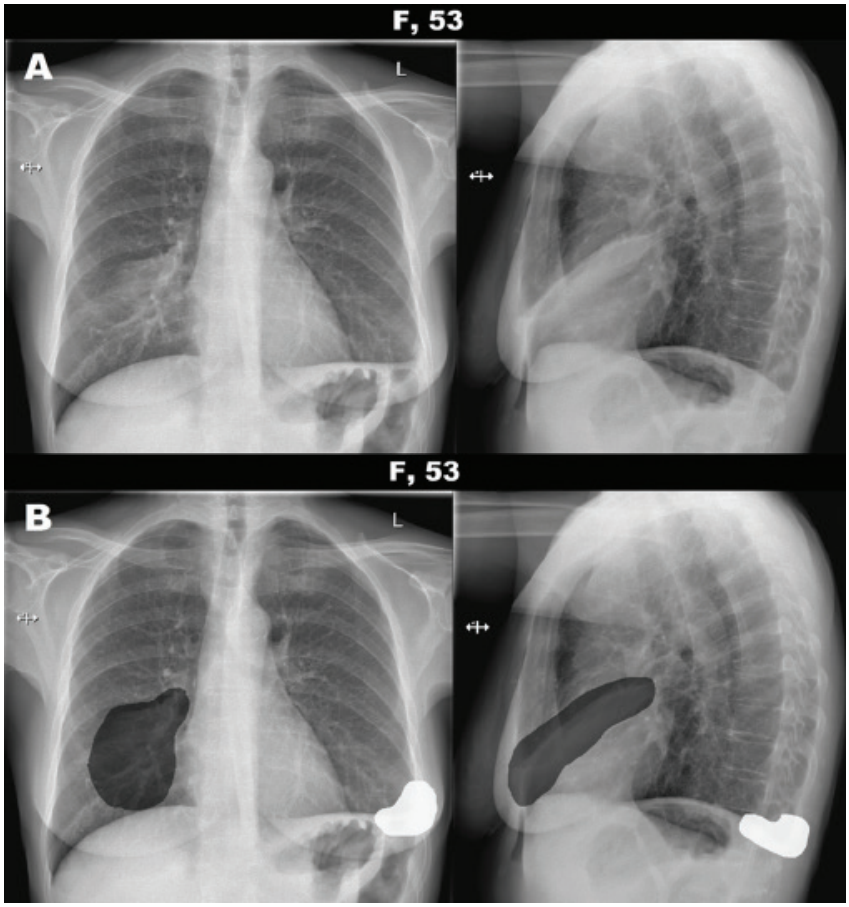
METHODS

The study was approved by the ethical review board of the Dutch Association for Medical Education. A total of sixteen Dutch first-year radiology residents participated in this study. One resident was additionally included but dropped out of residency before the second session. Residents started with thoracic (n=11), plain radiology (n=3), or abdominal (n=2) internships. Residents were employed in either academic hospitals (n=11) or nonacademic hospitals (n=5).

Chest radiographs

Experimental cases originated from a departmental image bank of 384 anonymized CXRs (19). Cases consisted of a screen with an upright posteroanterior and lateral CXR combined with participants` age and sex (See Figure 1a). All 384 CXRs were independently reevaluated by LU and RS, with respectively 11 and 33 years of experience in (chest) radiology. LU and RS outlined abnormal regions on CXRs (See Figure 1b), used as the Regions of Interest (ROIs). Approximately 1º margins were added to ROIs to account for potential eye-tracking imprecisions. Abnormal CXRs (223 of 384 CXRs) contained 2.3 lesions on average (range: 1-8) (see Table 1). The remaining 161 CXRs were considered normal and therefore contained no ROIs. CXRs were randomly sampled without replacement per participant for all experimental sessions. Per session 61% of 20 CXRs were abnormal, which did not significantly differ between sessions, X^2 (10) = 16.1, p = .11.

*Figure 1. Example CXR (A) with corresponding Regions of Interests (B). **



* The black ROI concerns atelectasis while the white ROI concerns a small pleural effusion. Note that abnormal findings may be seen on both the PA and the lateral image but are combined to one ROI in the analyses. This CXR, therefore, contains two ROIs.

Table 1. Abnormal findings and number of occurrences in the departmental image bank consisting of 384 CXRs.

Abnormal findings	Number of occurrences*
Added medical structures (cerclage wires, ECG leads, heart valve replacements, ICDs)	78
Aortic elongation	14
Atelectasis	51
Cardiomegaly	33

Table 1 Continues.

Abnormal findings	Number of occurrences*
Consolidation	47
Diaphragmatic herniation of the stomach	6
Enlarged hilus	17
Goiter	3
Interstitial disease	12
Mastectomy	5
Musculoskeletal (Fractures of clavicula, humerus, ribs, vertebra, scoliosis)	85
Normal variant (Collar rib, pectus excavatum, situs inversus)	10
Pleural calcification	8
Pleural effusion	25
Pneumoperitoneum	3
Pneumothorax (including subcutaneous emphysema)	18
Pulmonary masses	7
Pulmonary nodes	40
Redistribution	19
Signs of emphysema	65

*Note. CXRs can contain multiple abnormal findings

Apparatus

CXRs were shown on a 22" LCD screen (1650x1080 pixels) using Matlab 2015a with Psychtoolbox-3 (20). Eye movements were registered using an SMI RED250 remote eye tracker (SensoMotoric Instruments, Teltow, Germany) controlled using the SMITE toolbox (21). Head movements were not physically restricted. Fixations were classified using I2MC v1.1.1 (22) with default parameters, although adjacent fixations <40 pixels and 30 ms were merged, and fixations <80 ms were removed.

Procedure

The experimental sessions were held in weeks 2-4-6-8-11-15-20-26-43-42-51 of the first year of residency. Sessions were carried out individually at the participants' radiology department. Eye-tracker calibration was repeated until deviations <1.5º on x- and y-axes were acquired, or for a

maximum of five attempts. Participants were instructed to read CXRs as they would do in clinical practice. Participants left-clicked on any lesions they identified and pressed the spacebar to navigate to the next CXR. There were no time limits. The participants did not receive any feedback. The data was collected in the context of a larger experiment.

MEASURES

Reading time and visual search patterns

Reading time was measured in seconds. The total number of fixations, the average fixation duration, and the mean saccade length per CXR were used to measure visual search patterns. Per abnormal CXR, the proportion of the sum of fixation durations fixating on all ROIs combined divided by the total sum of fixation durations per image (proportion of abnormal dwell time) was calculated.

Lesion detection

Sensitivity and specificity were used as the lesion detection measures. Sensitivity was defined as the number of ROIs a participant clicked on, divided by the total number of ROIs per abnormal CXR. The sensitivity measure is not applicable to normal CXRs. Specificity was defined as normal areas where participants did not click. Specificity was 1 when participants did not click on normal areas and 0 when they clicked on normal areas. Specificity was calculated for all CXRs.

Analysis

Reading time and visual search patterns

Data inspection suggested that these data follow Weibull distributions (23), indicating that the variables are products of occurrences of events or hazards. Weibull distributions are usually found in survival analyses (24). Thus mixed-effect Cox survival analyses were used to assess longitudinal development of reading time and eye-tracking measures. Initial models consisted of the fixed factors time (months since the start of residency training), hospital category (academic versus nonacademic), CXR case number within sessions (ranging 1-20), and -if applicable- CXR category (abnormal/normal) and participants' individual intercept and slope on time as random factors.

Lesion detection

The number of lesions per abnormal CXR varied 1-8, and the sensitivity data were not normally distributed: The sensitivity values were concentrated near 0, .5, and 1. Moreover, specificity was bimodal per CXR; 0, or 1. Mixed-effects binary logistic regression models were thus used to assess the development of sensitivity and specificity. Two sensitivity analyses were performed: For the first sensitivity analysis, all scores >0 were coded 1, quantifying whether participants detected any abnormality. For the second sensitivity analysis, all scores <1 were coded 0, quantifying whether participants detected all abnormalities. One logistic analysis was performed for the already binary specificity measure. The initial models consisted of the factors time, hospital category, CXR case number, reading time, and -if applicable- CXR category as fixed factors, with CXRs as random factors. The reading times were first log-transformed to obtain normally distributed data.

Data modelling

Factors with the highest p-value were stepwise excluded from models. Chi-square tests on models' Akaike information criteria were used to assess the fit of different models (25). The modeling ended when new models did not have significantly better fits. R 3.6.3 (26) with the packages Coxme 2.2-16 (27) and Lme4 1.1-21 (28) were used for the analyses.

RESULTS

Demographics

The participants' mean age was 28.5 years (*SD* = 2.0) at the start of residency training, and 56% were male. Before the start of residency training, 12 of the 16 residents completed one or more elective radiology clerkships during medical school. After medical school, three participants immediately started with residency training while four first worked on Ph.D. projects, seven residents worked as junior doctors, and two worked both on Ph.D. projects and junior doctors. The average work experience as junior doctors was 17.7 months (range: 6-42).

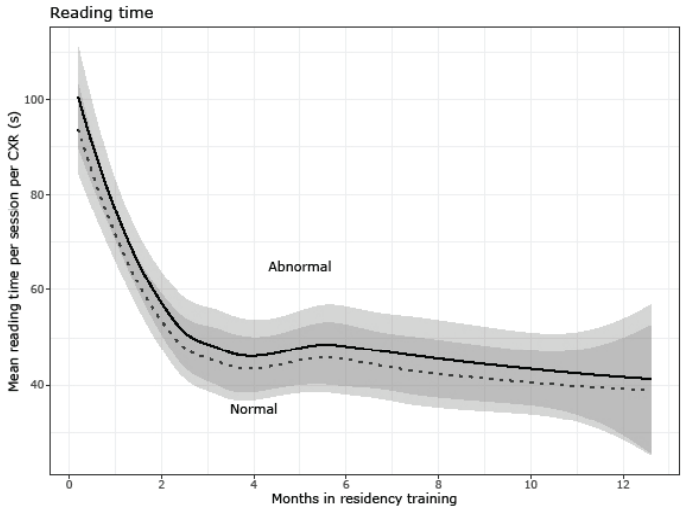
Missing values

In total, 169 of 176 experimental sessions were completed. Two participants missed two sessions, which were sessions 4 and 6. One participant missed session 1, and two participants missed session 7. Due to various reasons, such as software failure, failure to reach calibration data ≤1.5º, and substantial head movements of participants, eye-tracking data quality

could be insufficient. Eye-tracking data of 486 of 3290 CXRs (14%) had to be excluded due to insufficient data quality.

Reading time and visual search patterns
Reading time
The longitudinal development of reading time is visualized in Figure 2. The characteristics of the best-fitting model for reading time can be found in Table 2.

*Figure 2. Longitudinal development of reading time. **



*Reading time was averaged per session and CXR, and plotted over time in months since the start of residency training. The dotted line represents the reading time on normal CXRs and the solid line the reading time on abnormal CXRs. Grey areas surrounding the lines represent 95%-confidence intervals of the lines with darker grey areas indicating overlap of the 95%-confidence intervals of both lines.

Reading time decreased significantly over time, particularly during the first four months of training, when reading times were halved from 113 seconds to 56 seconds on average. This decrease continued yet was more gradual in the following eight months. A significant effect of the CXR category was found, indicating that reading times on abnormal CXRs were higher compared to normal CXRs. A small yet significant effect of the CXR case number was found, indicating that participants' reading time slightly decreased within the experimental sessions. The low random factor's variances of slopes indicated that the participants generally had similar longitudinal developments in reading time.

Table 2. Parameters of the final mixed-effects Cox regression model of Reading time.

Reading time (s)				
Fixed effects	*B*	exp(*B*)†	*SE*	*p*
Time in months	.13	1.14	.02	< .001**
CXR category (abnormal vs. normal)	-.29	.83	.04	< .001**
Case number within session (1 to 20)	1.02	1.02	.003	< .001**
Random effects	Variance	*SD*		
Intercept	.73	.85		
Slope	.004	.061		

B: regression coefficient, exp(*B*): exponential regression coefficient, *SE*: standard error of *B*, *SD*: standard deviation
† Note. Exponential regression coefficients <1 reflect an increase by the predictor variable while coefficients >1 reflect a decrease by the predictor variable on the outcome.

Visual search patterns

Visualizations of the number of fixations and the average fixation duration can be found in Figure 3. The characteristics of the final mixed-effect Cox regression analysis of the number of fixations and the average fixation duration can be found in Table 3.

*Figure 3. Longitudinal development of the number of fixations (A) and average fixation duration (B). **

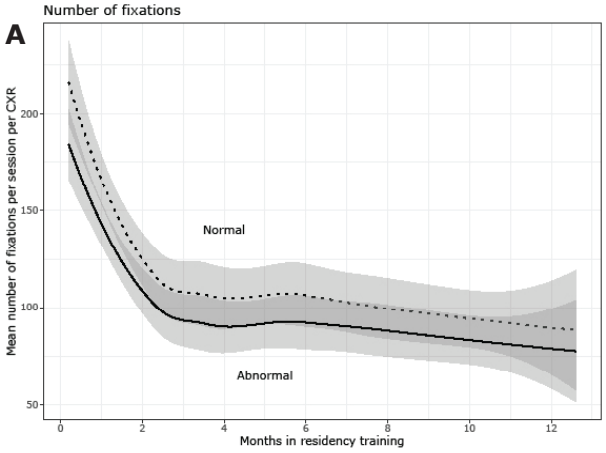
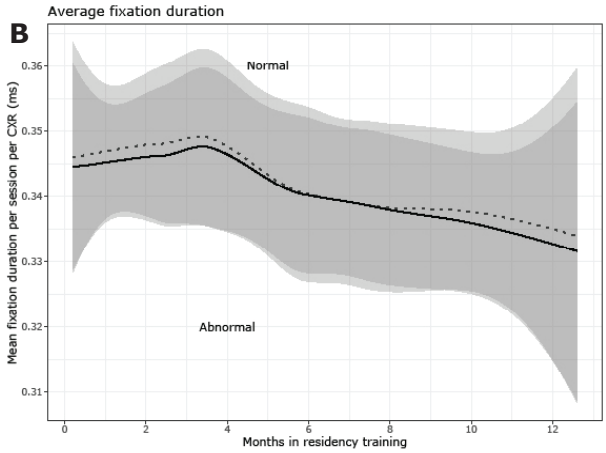


Figure 3 Continues.



* Number of fixations and average fixation duration plotted over time in months since the start of residency training, averaged per session, and CXR. Separate lines represent the separate patterns for normal (dotted lines) and abnormal (solid lines) CXR category. Grey areas surrounding the lines represent 95%-confidence intervals of the lines with darker grey areas indicating an overlap between the intervals.

The number of fixations significantly decreased over time, which parallels the longitudinal pattern of reading time. A significant effect of the CXR category was found: Participants fixated more often on the normal CXRs compared to the abnormal CXRs. The CXR case numbers had a significant yet small negative effect, indicating that participants gradually used fewer fixations per CXR within experimental sessions. Both the random factors intercept and slope had low variances, indicating minimal differences between participants' initial number of fixations and participants' longitudinal development.

The average fixation duration did not significantly change over time, although the *p*-value of .08 indicates a statistical trend of decreasing average fixation durations. The average fixation durations slightly increased in the first four months of training with a consecutive gradual decrease over the next eight months. No significant differences in the average fixation duration between normal and abnormal CXRs were found. Moreover, a small, yet significant negative effect of CXR order within sessions was found, indicative of a gradual decrease of the average fixation duration within sessions. The variance on the intercepts was quite high, indicating considerable differences between participants' average fixation durations.

Table 3. Parameters of the final mixed-effects Cox regression models for the number of fixations (3A) and average fixation duration (3B).

A.

Number of fixations

Fixed effects	*B*	exp(*B*)†	*SE*	*p*
Time in months	.11	1.12	.01	< .001**
CXR category	.17	1.18	.04	< .001**
Case number within session	.01	1.01	.003	< .001**
Random effects	Variance	*SD*		
Intercept	.35	.59		
Slope	.05	.002		

B.

Average fixation duration (s)

Fixed effects	*B*	exp(*B*)†	*SE*	*p*
Time in months	.030	1.03	0.02	.08
CXR category	-.030	.97	.004	.41
Case number within session	.022	1.02	.003	< .001**
Random effects	Variance	*SD*		
Intercept	.78	.89		
Slope	.004	.064		

B: regression coefficient, exp(*B*): exponential regression coefficient, *SE*: standard error of *B*, *SD*: standard deviation

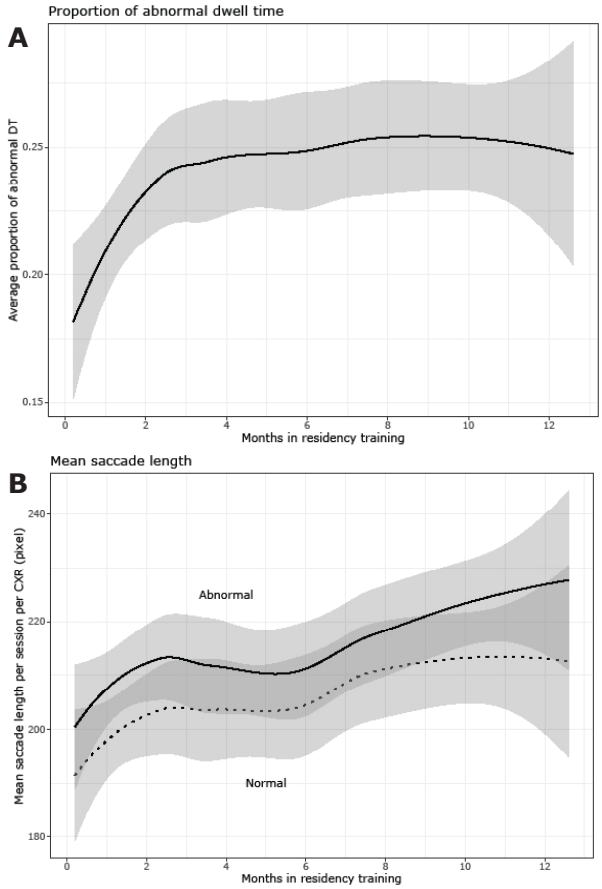
† Note. Exponential regression coefficients <1 reflect an increase by eye-tracking measure while coefficients >1 reflect a decrease by the eye-tracking measure on the outcome variable. The exponential coefficients indicate proportionate changes; i.e., for every month of training, the number of fixations decreases by 12% on average.

* *p* value <.05, ** *p* value <.01

The visualizations of the development of the proportion of abnormal dwell time and mean saccade length can be found in Figure 4. The characteristics of the final mixed-effect Cox regression analysis of the proportion of abnormal dwell time and mean saccade length be found in Table 4.

The proportion of abnormal dwell time significantly increased over time, indicating that participants increasingly fixated on abnormal instead of

*Figure 4. Longitudinal development of the proportion of abnormal dwell time (A) and mean saccade length (B). **



* The proportion of abnormal dwell time (DT) and mean saccade length plotted over time in months since the start of residency training, averaged per session, and CXR. Grey areas surrounding the lines represent 95%-confidence intervals of the lines. Concerning Figure 4B: Separate lines represent the separate patterns for normal (dotted lines) and abnormal (solid lines) CXR category. The darker grey area indicates an overlap between the 95%-confidence intervals of both lines.

normal areas throughout the first residency year. These changes were most pronounced during the first four months of training. No significant effect of case numbers on the proportion of abnormal dwell time within sessions was found.

The variance on the random factors was low, indicating minimal differences between participants' initial abnormal dwell time and similar participants' longitudinal developmental trajectories.

Table 4. Parameters of the final mixed-effects Cox regression models for the proportion of abnormal dwell time (4A) and mean saccade length (4B).

A.

Proportion of abnormal dwell time

Fixed effects	*B*	exp(*B*)†	*SE*	*p*
Time in months	-.021	.98	.01	.01*
Case number within session	.004	1.00	.004	.38
Random effects	**Variance**	***SD***		
Intercept	.0012	.11		
Slope	5.1e-3	2.6e-5		

B.

Mean saccade length (pixel)

Fixd effects	*B*	exp(*B*)†	*SE*	*p*
Time in months	-.056	.95	.02	< .001**
CXR category	-.36	.70	.04	< .001**
Case number within session	-.02	.98	.02	< .001**
Random effects	**Variance**	***SD***		
Intercept	.49	.70		
Slope	.005	.067		

B: regression coefficient, exp(*B*): exponential regression coefficient, *SE*: standard error of *B*, *SD*: standard deviation

† Note. Exponential regression coefficients <1 reflect an increase by eye-tracking measure while coefficients >1 reflect a decrease by the eye-tracking measure on the outcome variable. The exponential coefficients indicate proportionate changes; i.e. for every month of training the mean saccade length increases by 5% on average.

* *p* value <.05, ** *p* value <.01

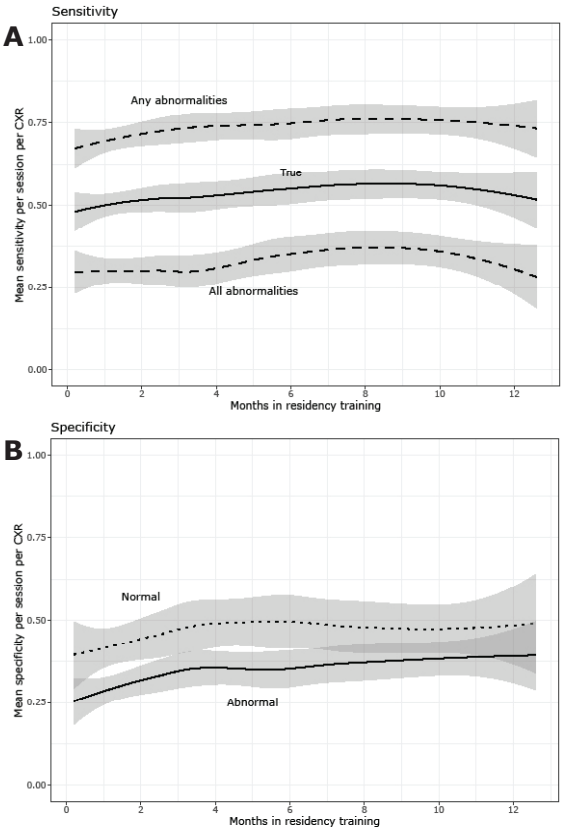
Mean saccade length significantly increased during the first year of residency training. This increase in saccade lengths was most pronounced during the first three months of training, followed by a plateau phase of two months and a further gradual increase during the last seven months. Furthermore, the mean saccade length was significantly longer on abnormal CXRs compared to normal CXRs. A small yet significant effect of case number was found:

participants' saccade lengths slightly increased throughout experimental sessions. The random factor' variance on the participants' slopes was low, indicating similar longitudinal developments of participants' mean saccade lengths.

Lesion detection

The visualizations of the longitudinal development of sensitivity and specificity can be found in Figure 5. Furthermore, the parameters of the final mixed-effects logistic regression models for sensitivity and specificity can be found in Table 5.

*Figure 5. Longitudinal development of sensitivity and specificity**



* Average sensitivity and specificity per session plotted over time in months since the start of residency training. Grey areas surrounding the lines represent 95%-confidence intervals of the lines with darker grey areas indicating overlap of the 95%-confidence intervals of both lines. 5A. Average (true) sensitivity is represented by the solid line. The probabilities to find any abnormalities and all abnormalities are also visualized (dashed lines). 5B. Separate lines are plotted for specificity on normal CXRs (dotted lines) and abnormal CXRs (solid lines).

Table 5. Parameters of the final mixed-effects logistic regression models for sensitivity (5A1 and 5A2) and specificity (5B).

A1

Sensitivity (any abnormalities)

Fixed effects	Odds‡	*95%-CI*	*p*
(intercept)	1.81	.59 - 5.62	.30
Time in months	1.03	.99 - 1.08	.11
Reading time (log)	1.31	1.01 - 1.70	.04*
Random effects	Variance	*SD*	
Case intercept	4.02	2.01	
Participant intercept †	.0070	.26	

A2

Sensitivity (all abnormalities)

Fixed effects	Odds‡	*95%-CI*	*p*
(intercept)	.13	.004 - .41	<.001**
Time in months	1.02	.98 - 1.06	.37
Reading time (log)	1.22	.94 - 1.57	.14
Random effects	Variance	*SD*	
Case intercept	3.98	1.99	
Participant intercept †	.15	.39	

B

Specificity

Fixed effects	Odds‡	*95%-CI*	*p*
(Intercept)	12.48	5.32 - 29.3	<.001**
Time in months	.98	.95 - 1.02	.36
CXR (abnormal vs. normal)	.54	.41 - .71	<.001**
Reading time (log)	.49	.41 - .58	<.001**
Random effects	Variance	*SD*	
Case intercept	1.03	1.02	
Participant Intercept	.22	.47	
Participant Slope	.0027	.005	

95%-CI: 95%-Confidence interval, *SD*: Standard deviation
‡ Note. Odds < 1 reflect lower probability on the outcome measure while odds >1 reflect a higher probability.
† Note. Since the sensitivity-models using both intercept and slope random factors for participants failed to converge
* $p < .05$, ** $p < .001$

On sensitivity, a gradual increase during the first nine months was found, followed by a slight decrease during the last three months. The first sensitivity analysis, odds to find any abnormalities, revealed a significant increase in odds for reading time, indicating that with longer reading times, the odds to find any abnormalities increased. The odds to find any abnormalities were not significantly different over time; however, the factor months in residency training needed to be included in the model for optimal fit. The second sensitivity analysis, odds to find all abnormalities, did not reveal a significant difference in odds for the factors reading time nor time since the start of residency training. The factor time again needed to be included in the model for optimal fit. Overall, the sensitivity analyses imply that the odds to detect any abnormalities increased with longer reading times but that longer reading times do not lead to the detection of all abnormalities. Moreover, the sensitivity analyses could not reveal a significant change over time since the start of residency training, although reading time needed to be included for optimal fit. Moreover, the variance was high for the random factor cases, indicating that participants' sensitivity differed considerably between cases when compared to the other random factors.

Specificity gradually increased during the first six months, followed by a plateau phase during the next six months. The specificity' odds significantly decreased with longer reading times. Moreover, specificity was significantly higher for normal CXRs compared to abnormal CXRs. The factor months in residency training did not have a significant, independent effect on specificity odds but again needed to be included in the model for an optimal fit. Moreover, the variance was high for the random factor case, and specificity thus differed considerably between the cases.

DISCUSSION

This longitudinal prospective eye-tracking study addressed the following research questions:

1. How do reading times and visual search patterns change over time in the first year of residency?

2. How does lesion detection change over time in the first year of residency?

First, in only four months, reading times were halved, accompanied by the development of more efficient visual search patterns. The decrease in

reading times is reflected in the decrease in the total number of fixations as fixations occur 2-3 times per second (12). Furthermore, the sharp increase in the proportion of abnormal dwell time during the first four months indicates that residents learn to focus on abnormal areas, while longer mean saccades and the longitudinal trend on lower average fixation durations over 12 months are also indicative of more efficient visual search (6).

When comparing visual search patterns on abnormal and normal images, higher reading times, longer saccade lengths, and lower number of fixations were found on the abnormal images. Differences in reading times and visual search parameters on abnormal and normal images have been previously found (4). Our findings indicate that residents quickly learn to adapt their visual search to image characteristics. Additionally, this also indicates that the longitudinal development of visual search patterns found in this study may not be unconditionally extrapolated to other image types. The characteristics of the specific image type should thus be taken into account when studying visual search parameters over time (29).

Second, as residents` visual search patterns are increasingly efficient, are residents also becoming increasingly effective in lesion detection? The sensitivity and specificity analyses showed that residents' lesion detection remained constant over time, yet detection was significantly influenced by reading times. Reading times change over time, and any longitudinal effects on lesion detection are probably indirect effects. Additionally, residents' specificity on abnormal images was higher compared to normal images. Many cardiopulmonary diseases have multiple abnormal findings on CXRs, and the identification of one abnormality may shift residents` criterion to call normal tissue abnormal (19).

Overall, reading time, visual search patterns, and lesion detection particularly changed during the first four months of residency training. By investigating the reading process, a part of the residents' learning process was uncovered (9, 30). Monitoring reading times and visual search patterns could thus help to tailor residency training to individual needs. When reading times and visual search parameters are still dramatically changing, the resident is probably still developing visual search and may need more exposure to a specific image type. Measuring visual search patterns could thus be used to individualize residency training, but further research on individual training adjustments based on changes in reading processes is advised.

Moreover, measuring visual search patterns can have added value to residents' competency assessment. On the one hand, summative tests are commonly aggregated over all radiology subspecialties (8) and can be too generic. On the other hand, competency assessment in the workplace generally consists of only little cases and may be too specific (31). Moreover, it may be challenging to intentionally change visual search patterns (12, 32) while summative test scores can be intentionally inflated by, e.g., studying previous assessments (33). Therefore, measuring visual search patterns over a series of cases could be used as an objective assessment on one image type and provide in-depth information about the development of residents (11). However, which visual search parameters are most suitable for competency assessment has yet to be investigated. The eye-tracking parameters of this study showed different developmental trajectories, and some, e.g., the average fixation duration, may be less suited to assess residents' competency development.

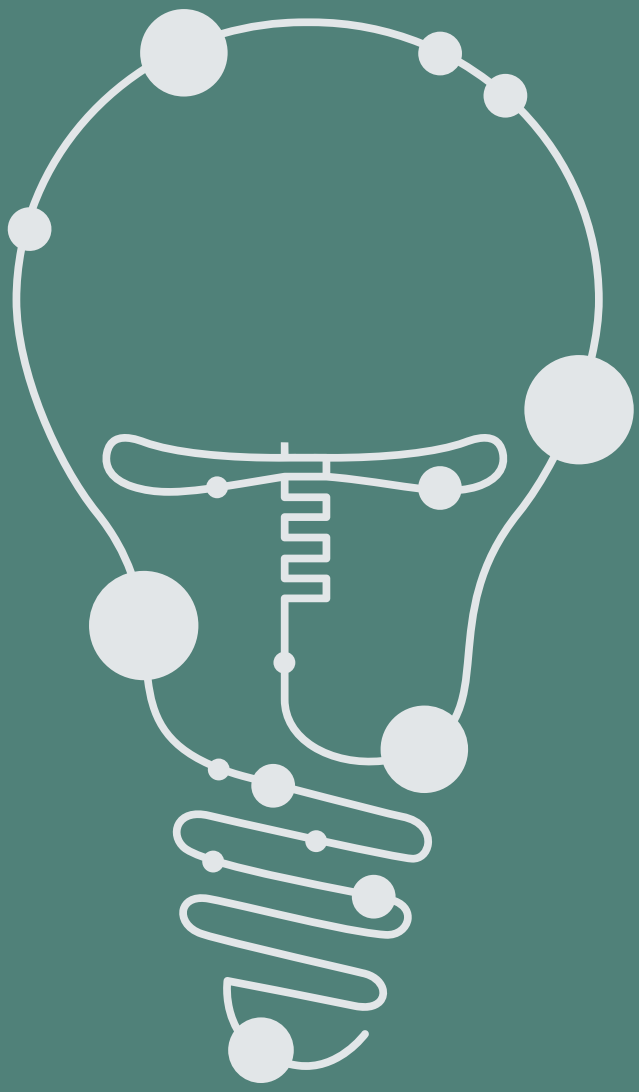
This study has some limitations. First, no specific clinical information was provided. Clinical information may particularly affect novices' visual search and lesion detection as novices are learning to discern abnormal from normal features, whereas experienced readers seem to be less influenced by the provision of clinical information (34). Investigating potential interactions between clinical information and expertise level was not the purpose of this study, and clinical information was thus not provided. Moreover, in clinical practice, often only little information is provided. Second, this study only focused on CXRs since most residents start learning to read CXRs at the beginning of residency. However, a substantial minority completed their thoracic internship later on, and all residents probably had continued exposure to CXRs as they started on-call shifts. Indeed, visual search parameters, such as the average fixation duration and the mean saccade length, continued to change over 12 months.

In conclusion, residents' reading time is halved during the first four months of residency training when reading CXRs. This development is accompanied by more efficient visual search patterns, while lesion detection remained stable. Reading times and visual search patterns slightly differed between abnormal and normal CXRs, which indicate that residents also learn to adapt their visual search to image characteristics. Monitoring visual search patterns can provide additional insights on residents' development and may be used to tailor residency training to individual needs.

REFERENCES

1. Krupinski EA. Current perspectives in medical image perception. Attention Perception & Psychophysics. 2010;72(5):1205-17.
2. van der Gijp A, Ravesloot CJ, Jarodzka H, van der Schaaf MF, van der Schaaf IC, van Schaik JPJ, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education. 2017;22(3):765-87.
3. Gijp A, Schaaf MF, Schaaf IC, Huige JCBM, Ravesloot CJ, Schaik JPJ, et al. Interpretation of radiological images: towards a framework of knowledge and skills. Advances in Health Sciences Education. 2014:1-16.
4. Kok EM, de Bruin ABH, Robben SGF, van Merriënboer JJG. Looking in the Same Manner but Seeing it Differently: Bottom-up and Expertise Effects in Radiology. Applied Cognitive Psychology. 2012;26(6):854-62.
5. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department: is the radiologist really necessary? Postgraduate medical journal. 2003;79(930):214-7.
6. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. Educational Psychology Review. 2011:1-30.
7. Bertram R, Kaakinen J, Bensch F, Helle L, Lantto E, Niemi P, et al. Eye Movements of Radiologists Reflect Expertise in CT Study Interpretation: A Potential Tool to Measure Resident Development. Radiology. 2016;281(3):805-15.
8. Ravesloot CJ, van der Schaaf MF, Kruitwagen C, van der Gijp A, Rutgers DR, Haaring C, et al. Predictors of Knowledge and Image Interpretation Skill Development in Radiology Residents. Radiology. 2017;284(3):758-65.
9. Gegenfurtner A, Kok E, van Geel K, de Bruin A, Jarodzka H, Szulewski A, et al. The challenges of studying visual expertise in medical image diagnosis. Med Educ. 2017;51(1):97-104.
10. Wu CC, Wolfe JM. Eye Movements in Medical Image Perception: A Selective Review of Past, Present and Future. Vision (Basel, Switzerland). 2019;3(2).
11. Kok EM. Eye tracking: the silver bullet of competency assessment in medical image interpretation? Perspectives on medical education. 2019;8(2):63-4.
12. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Weijer J. Eye Tracking: A Comprehensive Guide to Methods and Measures: Oxford University Press; 2011.
13. van Geel K, Kok EM, Dijkstra J, Robben SG, van Merrienboer JJ. Teaching Systematic Viewing to Final-Year Medical Students Improves Systematicity but Not Coverage or Detection of Radiologic Abnormalities. Journal of the American College of Radiology : JACR. 2017;14(2):235-41.
14. Hessels RS, Niehorster DC, Nyström M, Andersson R, Hooge ITC. Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. Royal Society Open Science. 2018;5(8):180502.
15. Nodine C, Mello-Thoms C. The role of expertise in radiologic image interpretation. In: Samei E, Krupinski E, editors. The Handbook of Medical Image Perception and Techniques. Cambridge: Cambridge University Press; 2010. p. 139-56.
16. Kundel HL, Nodine CF. A visual concept shapes image perception. Radiology. 1983;146(2):363-8.
17. Haider H, Frensch PA. Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. Journal of Experimental Psychology-Learning Memory and Cognition. 1999;25(1):172-90.
18. Pusic M, Pecaric M, Boutis K. How much practice is enough? Using learning curves to assess the deliberate practice of radiograph interpretation. Acad Med. 2011;86(6):731-6.
19. Geel Kv, Kok EM, Aldekhayel AD, Robben SGF, van Merriënboer JJG. Chest X-ray evaluation training: impact of normal and abnormal image ratio and instructional sequence. Medical Education. 2019;53(2):153-64.
20. Kleiner M, Brainard D, Pelli D. What's new in Psychtoolbox-3? 2007.
21. Niehorster DC, Nyström M. SMITE: A toolbox for creating Psychophysics Toolbox and PsychoPy experiments with SMI eye trackers. Behavior Research Methods. 2020;52(1):295-304.

22. Hessels RS, Niehorster DC, Kemner C, Hooge ITC. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). Behavior Research Methods. 2017;49(5):1802-23.
23. Cox DRS, Oakes D. Analysis of survival data: London : Chapman and Hall; 1984.
24. Fox J. Cox proportional-hazards regression for survival data.
25. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19(6):716-23.
26. R Development Core Team R. R: A language and environment for statistical computing. R foundation for statistical computing Vienna, Austria; 2011.
27. Therneau TM. Mixed Effects Cox Models [R package coxme version 2.2-16]. 2018.
28. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. 2015. 2015;67(1):48.
29. Bertram R, Helle L, Kaakinen JK, Svedstrom E. The Effect of Expertise on Eye Movement Behaviour in Medical Image Perception. PloS one. 2013;8(6).
30. Kok EM, Jarodzka H. Before your very eyes: the value and limitations of eye tracking in medical education. Med Educ. 2017;51(1):114-22.
31. Durojaiye AB, Snyder E, Cohen M, Nagy P, Hong K, Johnson PT. Radiology Resident Assessment and Feedback Dashboard. RadioGraphics. 2018;38(5):1443-53.
32. Jarodzka H, Holmqvist K, Gruber H. Eye tracking in Educational Science: Theoretical frameworks and research agendas. Journal of Eye Movement Research. 2017;10(1).
33. Moravec T, Štěpánek P, Valenta P. The Impact of Progress Testing of Students on their Results at Final Exam. Procedia - Social and Behavioral Sciences. 2015;174:3702-6.
34. Cooper L, Gale A, Darker I, Toms A, Saada J. Radiology image perception and observer performance: How does expertise and clinical information alter interpretation? Stroke detection explored through eye-tracking. 2009.



Chapter 5

Reversal of the hanging protocol of Contrast-Enhanced Mammography leads to similar diagnostic performance yet decreased reading times

Koos van Geel, Ellen M. Kok, Jorian P. Krol, Ivo P.L. Houben, Ruud M. Pijnappel, Jeroen J.G. van Merriënboer, Marc B.I. Lobbes.

European Journal of Radiology. 2019;117:62-8.

ABSTRACT

Objectives

Contrast-enhanced mammography (CEM) was found superior to Full-Field Digital Mammography (FFDM) for breast cancer detection. Current hanging protocols show low-energy (LE, similar to FFDM) images first, followed by recombined (RC), post-contrast images. However, evidence regarding which hanging protocol leads to the most efficient reading process and the highest diagnostic performance is lacking. This study investigates the effects of hanging-protocol ordering on the reading process and diagnostic performance of breast radiologists using eye-tracking methodology. Furthermore, it investigated differences in reading processes and diagnostic performance between LE, RC, and FFDM images.

Materials and methods

Twenty-seven breast radiologists were randomized into three reading groups: LE–RC (commonly used hangings), RC-LE (reversed hangings), and FFDM. Thirty cases (nine malignant) were used. Fixation count, net dwell time, and time-to-first fixation on malignancies as visual search measures were registered by the eye-tracker. Reading time per image was measured. Participants clicked on suspicious lesions to determine sensitivity and specificity. Area-under-the-ROC-curve (AUC) values were calculated.

Results

RC-LE scored identical on visual search measures, $t(16) = -1.45$, $p = .17$ or higher p- values, decreased reading time with 31%, $t(16) = -2.20$, $p = .04$, while scoring similar diagnostic performance compared to LE-RC, $t(13.2) = -1.39$, $p = .20$ or higher p-values. The reading process was more efficient on RC compared to LE images. Diagnostic performance of CEM was superior to FFDM; $F(2, 26) = 16.1$, $p < .001$. Average reading time did not differ between the three groups, $F(2, 25) = 3.15$, $p = .06$.

Conclusion

The reversed CEM hanging protocol (RC-LE) scored similarly on diagnostic performance compared to LE-RC, while reading time was a third faster. Abnormalities were interpreted quicker on RC images. An RC-LE hanging protocol is therefore recommended for clinical practice and training. The diagnostic performance of CEM was (again) superior to FFDM.

INTRODUCTION

Contrast-enhanced mammography (CEM) has been shown to be superior to Full-Field Digital Mammography (FFDM) for both the detection of breast cancer and the evaluation of disease extent (1-4). A typical CEM exam consists of a low-energy (LE) image, which is comparable to FFDM (5, 6), and a post-contrast recombined (RC) image, which shows areas of contrast uptake (1). Prior studies found that all diagnostic parameters of CEM were significantly higher when compared to FFDM (2). It even matched the diagnostic accuracy of breast MRI, which is generally considered to be the most accurate breast imaging modality (7-9). At present, all vendors present CEM-cases on their workstations using a hanging protocol (the order in which the images are presented to the radiologist) showing the LE images first, followed by the RC images, either as overlay or as a separate image (3, 10, 11). However, evidence on what hanging protocol is most effective is lacking (12), as there is no knowledge on how radiologists read CEM exams in clinical practice.

Eye-tracking methodology allows us to investigate how the sequence of LE images and RC images of a hanging protocol affects the reading process. An eye-tracker measures where, when, and for how long a radiologist looks during the reading process (13). A particular image area, such as a lesion visible on one of the images, could draw the radiologist's attention, moving the eyes to this area. Eye movements thus reflect the radiologist's directed attention (14, 15). Eye-tracking can objectively measure whether radiologists find lesions faster and fixate longer on lesions in a certain hanging protocol.

This paper aims to investigate the effects of hanging protocols on the reading process and diagnostic performance of breast radiologists. Also, it aims to investigate the differences in the reading process and diagnostic performance between CEM and FFDM. We hypothesize that: (1) participants of a (reversed) RC-LE hanging protocol will be more efficient and will score higher on diagnostic performance in the reading of cases compared to participants using the (regular) LE-RC hanging protocol; (2) participants will be more efficient in their reading process and will score higher on the diagnostic performance of RC images compared to LE images; (3) participants using any CEM protocol will be more efficient and score higher on diagnostic performance compared to participants using conventional FFDM.

MATERIAL AND METHODS

Participants

Breast radiologists, fellows in breast radiology, and residents in an advanced rotation of breast radiology were eligible for participation in this experiment. To acquire a diverse set of radiologists, members of the European Society of Breast Imaging (EUSOBI) were invited by e-mail to participate. To accommodate the participation of members from abroad, part of the data collection took place in a dedicated room on-campus during the 2018 congress of the European Society of Radiology (ESR).

Participants were randomly allocated to one of three experimental groups; 1. the FFDM group, which evaluated the FFDM case only; 2. the LE-RC group, which evaluated the LE image first, followed by the RC image of each CEM-case, similar to current used hanging protocols; 3. and the RC-LE group, which evaluated the RC image first followed by the LE-image of each CEM-case (i.e., the 'reversed' hanging protocol).

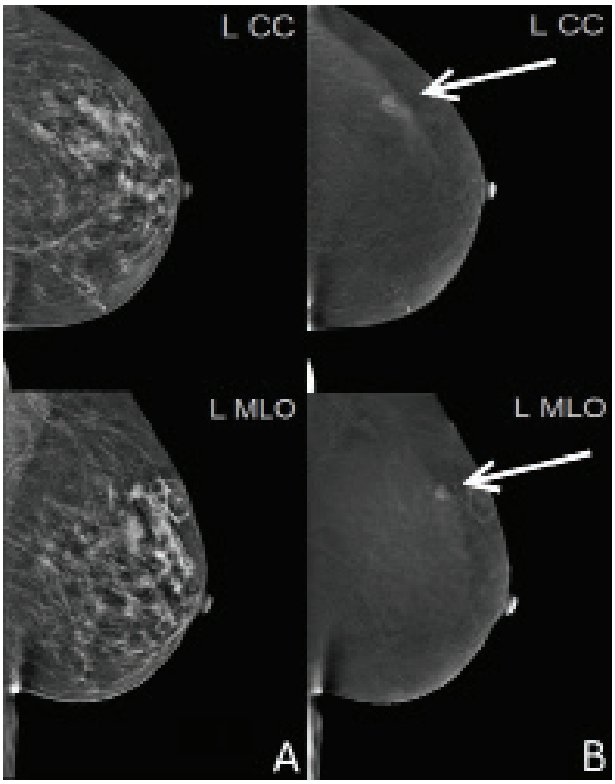
MATERIALS

Images

For this study, the CEM-cases of 30 patients were used. As all patient cases were anonymized, our certified ethical committee waived the need to obtain informed consent from patients. The images originated from our hospital's database consisting of CEM-cases acquired between 2012 and 2016 (4). The CEM principle and its imaging protocol were described earlier (3, 16). In summary, an LE and a high-energy (HE) image are obtained of both breasts in the standard mediolateral oblique (MLO) and craniocaudal (CC) views two minutes after intravenous administration of 1.5 mL/kg body weight iodine-based contrast medium (Iopromide 300 mg/ml) with a flow rate of 3 mL/s followed by a saline flush. The LE and HE images are recombined to create the RC image, which visualizes areas of contrast uptake.

Due to the resolution of the non-diagnostic LCD screen (specifications regarding the reading set-up are delineated under Apparatus), the images of only the right or left breast were used in the experiment. LE images and RC images of a patient case were shown after each other during the experiment, whereas sequence (RC-LE or LE-RC) varied as described before. A typical CEM exam used in the experiment is visualized in Figure 1.

*Figure 1. Typical example of a CEM exam used in the experiment, showing only the left breast.**



* On the left (A), CC and MLO views of the low-energy (LE) images are shown, which are similar to FFDM. On the right (B), CC and MLO views of the recombined (RC) images are shown. An area of contrast uptake can be seen (arrows), suspicious for breast cancer. Biopsy results confirmed invasive breast cancer at this site. A and B were shown after each other during the experiment.

Out of all CEM-cases, nine cases were selected that were considered typical malignant cases selected by a breast radiologist with four years of experience in CEM. Each malignant case contained one malignant lesion. All (histologically proven) malignancies were invasive carcinomas of no special type (NST). The size of the abnormalities on the images ranged from 0.9x0.6 cm (31x21 pixels) to 1.8x1.4 cm (65x50 pixels). Additionally, 21 CEM-cases with only negative findings were selected. Images that contained benign lesions such as simple cysts or fibroadenomas, artifacts (17), or (micro)calcifications were excluded. The malignant-benign ratio of

the selected cases was similar to the percentage malignancies of our CEM database, which is 28% (4). There were four cases with a breast density category A, twelve cases with a B category, ten cases with a C category, and three cases were considered to have a breast density category D.

Apparatus

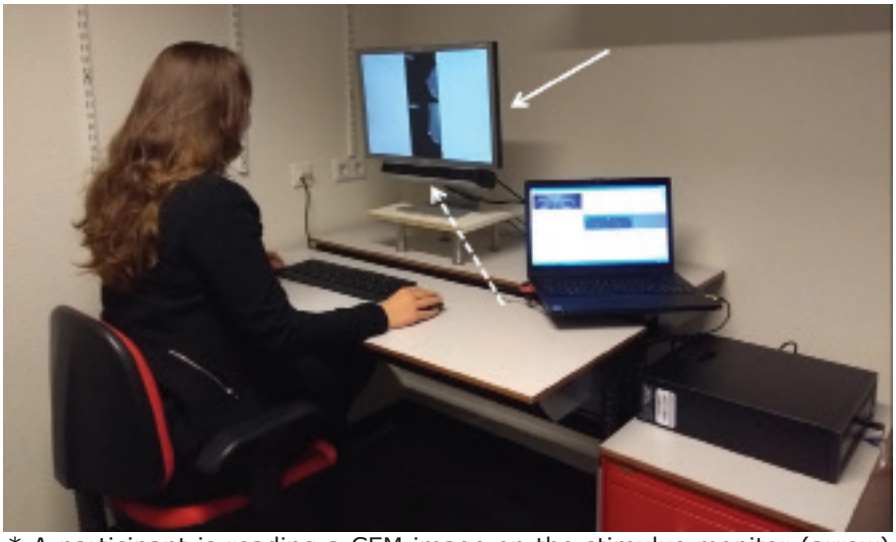
Eye movements were measured using a SensoMotoric Instruments (SMI, Teltow, Germany) 250 Hz remote eye tracker. Participants` head movements were not physically restricted, although they were instructed to avoid head movements as much as possible. As the right eye is generally dominant, the eye movements of the participants' right eyes were used (13). The stimuli were shown on a Dell 22" liquid screen display with a resolution of 1080x1650 pixels in portrait set-up. The distance between the participant and monitor was approximately 70 centimeters, and the visual angle θ is thus 41º. A dispersion-based detection algorithm was used, and the minimal fixation duration was set to 22 ms. The eye-tracker was calibrated using a five-point calibration prior to and halfway through the experiment. Calibration was repeated until a deviation small than 1º of visual angle on the x- and the y-axis was obtained. Eye-tracking data of one participant was excluded from the analysis as the eye-tracking deviations were greater than 1.0º visual angle.

Participants used the computer of the experiment computer to click on any lesions suspicious for malignancy they identified on an image, and the space bar to navigate to the next image or case. The experimental set-up is visualized in Figure 2.

PROCEDURE

The experiment was carried out by each participant individually. Participants were first instructed that they were going to evaluate 30 patient cases of women recalled from a breast cancer screening program. They were instructed to search for malignant masses but were informed that not all images contained masses. Participants were instructed to click on all areas which they deemed malignant (BIRADS 4-5) on the MLO- as well as CC-part of every image; participants of the FFDM group were instructed to click on the FFDM image, and participants of the CEM groups were instructed to click on the LE as well as the RC image. They were informed that the images would not contain any technical artifacts, architectural disturbances, (micro)calcifications, or benign masses. When participants of the CEM

*Figure 2. Photo of the experimental set-up.**



* A participant is reading a CEM image on the stimulus monitor (arrow) while the eye movements are registered by the eye-tracker (dashed arrow).

groups finished the reading of the first MLO- and CC-images of a patient case (respectively LE-images or RC-images), they should press the space bar for the second MLO- and CC-images. It was not possible to return to the previous images of a patient case and to reevaluate the previous images.

After receiving the instructions, the participants subsequently wrote down their age, sex, hospital where they were employed, number of years licensed as a radiologist, number of workdays per week working as a breast radiologist, fellow or resident, and, if applicable, number of years of experience with evaluating CEM images. Participants then evaluated a practice case to check whether all instructions were clear, followed by the five-point calibration procedure for the eye tracker where after they started with the reading of the patient cases. After 15 patient cases, participants had a short break of two minutes, followed by a recalibration before they continued with the last 15 patient cases. Participants received no feedback on their performance throughout the experiment.

ANALYSIS

Reading process measures

To investigate the efficiency of the reading process, the following eye-tracking measures were used: Average fixation count, average net dwell time, average time-to-first fixation, and average reading time. Average fixation count was defined as the average number the participants` eyes stood still (fixated) on malignant areas, which were the areas of interest (AOIs), averaged over images (either FFDM/LE or RC). Average net dwell time was defined as the total time participants fixated on the AOIs, averaged over images (either FFDM/LE or RC) (13). Average time-to-first fixation was defined as the time the eyes first fixated on a malignant area, averaged over images (either FFDM/LE or RC). Furthermore, average reading time was defined as the total time that participants needed to evaluate an image, and a case: Average reading time was calculated per image (either FFDM/LE or RC image) and per case (FFDM/LE and RC image reading time combined).

Diagnostic performance

Sensitivity and specificity were used as measures of diagnostic performance. Sensitivity was defined as the number of malignant areas a participant clicked on, divided by the total number of nine malignant areas of the experiment. Specificity was defined as the number of images where a participant did not click on benign areas, divided by the total number of 30 images of the experiment. The mouse clicks on the last image (LE or RC for the two CEM groups and FFDM for the FFDM group) of each of the 30 patient cases participants evaluated were used to calculate sensitivity and specificity. To analyze differences in diagnostic performance on separate LE and RC images of a patient case, sensitivity per image category, and specificity per image category were calculated.

Furthermore, the participants worked in different hospitals and potentially had different criteria to call a lesion malignant. "Over"calling would result in high sensitivity yet low specificity. Therefore, an aggregate measure of diagnostic performance was necessary. Participants` sensitivity and specificity were used to calculate individual receiver-operator characteristic (ROC) curves, and participants` value of area under the ROC curve (AUC) was used as an aggregate measure of diagnostic performance.

Statistics

To compare the RC-LE group with the LE-RC group (hypothesis 1), independent *t*-tests were used with average fixation count, average net dwell time, average time-to-first fixation, and average reading time as dependent variables of the reading process and sensitivity, specificity, and AUC-values as dependent variables of diagnostic performance. To compare participants' reading process (average fixation count, average net dwell time, average time-to-first fixation, and average reading time) on RC images and LE images of the same patient cases (hypothesis 2), paired sample *t*-tests were used. Finally, to compare the reading process (average net dwell time, average time-to-first fixation, and average evaluation time), and diagnostic performance (sensitivity, specificity, and AUC-value) of the two CEM-groups versus the FFDM group (hypothesis 3) one-way analyses of variance (ANOVAs) were used with CEM-groups contrasted to the FFDM group. Data analysis was performed using SPSS (IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp. Released 2016).

RESULTS

Demography of participants

The participants (*n* = 27, mean age = 42.1 years, Standard Deviation (*SD)* = 10.4, 67% female) originated from 18 different hospitals in seven European countries (the Netherlands, Belgium, United Kingdom, France, Ireland, Italy, and Spain). Two participants were residents in an advanced rotation of breast radiology, and the other 25 were radiologists with an average career span of 10 years, *SD* = 9.0). The participants worked as breast radiologists for an average of 3.9 days per week (*SD* = 1.2). Nine participants also had prior experience with evaluating CEM (mean span = 3.15 years, *SD* = 1.62). The three study groups did not significantly differ in any of the following demographic factors: age, gender, country, career span, experience with CEM, and average breast radiology working hours per week; in one-way ANOVAs comparing the three groups, the highest F-value was $F(2, 26)= 1.91$, $p = .19$.

The reading process and diagnostic performance measures per group can be found in Table 1. The independent t-tests on the reading process and diagnostic performance measures can be found in Table 2.

Table 1. Reading process and diagnostic performance measures per group and image category.

		RC-LE (N=10)		LE-RC (N=9)		FFDM (N=8)
		RC	LE	RC	LE	/ FFDM
Reading process measures	Unit	μ (*SD*)	μ (*SD*)	μ (*SD*)	μ (*SD*)	/ μ (*SD*)
Average fixation count	#	3.24 (1.24)	3.58 (1.74)	2.75 (1.74)	4.15 (2.23)	5.26 (2.94)
Average net dwell time	ms	1280 (356)	1301 (495)	1025 (502)	1444 (780)	2243 (1248)
Average time-to-first fixation	ms	773 (333)	888 (481)	1107 (635)	1295 (691)	1197 (580)
Average reading time per image	s	6.30 (2.44)	7.31 (4.07)	5.82 (2.76)	14.0 (4.26)	12.8 (7.19)
Diagnostic performance measures	Unit	μ (*SD*)	μ (*SD*)	μ (*SD*)	μ (*SD*)	/ μ (*SD*)
Sensitivity overall	%	100 (0)		98 (7.4)		79 (13)
Sensitivity per image	%	99 (4.0)	100 (0)	98 (7.0)	80 (11)	
Specificity overall	%	87 (15)		94 (7.2)		80 (17)
Specificity per image	%	90 (10)	87 (15)	94 (7.2)	80 (8.4)	
Area under the curve		.93 (.075)		.96 (.048)		.80 (.055)

Influence of hanging protocol

The reading process and diagnostic performance measures of the two CEM groups are found in the first two columns of Table 1. The eye movement measures of the two CEM groups are generally similar. These findings indicate a similar reading process for the participants of the two CEM

groups. Furthermore, the RC-LE group took 6.3 seconds to read the RC image and 7.3 to read the LE image, while the LE-RC group respectively took an average of 5.8 and 14 seconds to read the RC and LE images. Finally, the diagnostic performance measures were similar between the two CEM groups with an AUC value of .93 for the RC-LE group and .96 for the LE-RC group.

Furthermore, the results of the independent *t*-tests concerning hypothesis 1 are found in Table 2. The t-tests on the eye movement measures are all nonsignificant. Thus participants of both the RC-LE and LE-RC groups had similar eye movement measures on the RC images and LE images. However, a significant effect of the hanging protocol was found on reading time per case, indicating that participants of the RC-LE group needed 6.20 seconds less to evaluate a patient case. The participants of both groups needed a similar amount of time to evaluate the RC images. A significant effect of the hanging protocol on the reading time of the LE-images was found, indicating that participants of the RC-LE group needed less time to evaluate the LE-image compared to the LE-RC group. Furthermore, the t-tests on diagnostic performance were nonsignificant, indicating a similar sensitivity, specificity, and AUC-value for both of the CEM-groups.

Table 2. Independent t-tests on the reading process and diagnostic performance measures with RC-LE versus LE-RC group as independent variables.

Reading process measure	*t*	*df*	*p*	Mean difference (*95%-CI*)
Fixation count (RC)	.68	16	.51	.48 (-1.02 - 1.99)
Fixation count (LE)	-.58	16	.57	-.56 (-2.61 - 1.49)
Net dwell time (RC)	1.24	16	.23	205 (-180 - 689)
Net dwell time (LE)	-.47	16	.65	308 (-796 - 509)
Time-to-first fixation (RC)	-1.40	16	.18	-334 (-841 - 173)
Time-to-first fixation (LE)	-1.45	16	.17	-407 (-1008 - 194)
Reading time per case	-2.20	16	.04	-6.2 (-12.2 - 2.65)
Reading time per image (RC)	.49	16	.63	5.98 (-2.00 - 3.20)

Table 2 Continues.

Diagnostic performance measure	*t*	*df*	*p*	Mean difference (*95%-CI*)
Sensitivity overall	1.00	8.00	.35	.25 (-.032 - .082)
Sensitivity (RC)	.52	17	.61	.026 (-.042 - .069)
Sensitivity (LE)	5.49	8.00	.001	.20 (.13 - .27)
Specificity overall	-1.39	13.17	.20	.053 (-.19 - .041)
Specificity (RC)	-.94	17	.36	-.040 (-.12 - .046)
Specificity (LE)	1.24	17	.23	.056 (-.050 - .19)
Area under the curve	-.83	17	.42	.030 (-.086 - .037)

Differences between RC images and LE images on the reading process and diagnostic performance measures

The reading process and diagnostic performance measures on the RC and LE images are found in the first four columns of Table 1. The average fixation count is lower on the RC images; 3.24 and 2.75 fixations on the RC images and respectively 3.58 and 4.15 fixations on the LE images. This difference is significant with a p-value of .003. The average net dwell time is also lower on RC images, 1280 and 1025 ms compared to 1301 and 1444 ms for the LE images. This difference is also significant, $p = .029$.

Time-to-first fixation is lower on RC images, 773 and 1107 ms compared to 888 and 1295 ms for the LE images, yet this difference is nonsignificant, $p = .22$. Furthermore, participants read RC images faster compared to LE-images, $p < .001$. Finally, participants scored higher on sensitivity on RC images, 99% and 98% compared to 100% and 80% on LE images, $p = .004$ and higher on specificity on RC images, 90% and 94% compared to 85% and 80% on LE images, $p = .019$ compared to LE-images. The complete results of the paired sample t-tests of RC compared to LE images are found in Table 3.

Comparison of CEM-groups to FFDM group

For hypothesis 3, no effect of group was found on fixation count on the LE and FFDM images, $F(2, 25) = 1.09$, $p = .35$, contrast $t(23) = -1.39$, $p = .18$. A tendency towards lower net dwell time for the CEM-groups was found, $F(2, 23) = 2.79$, $p = .08$, contrast $t(8.74) = -1.86$, $p = .10$. No group effect was found on average time-to-first fixation, $F(2, 23) = 1.16$,

Table 3. Paired sample t-tests of the reading process and diagnostic performance measures on RC and LE images.

Variable	*t*	*df*	*p*	Mean difference (*95%-CI*)
Fixation count	-3.47	17	.003	-.87 (-1.40 - -.33)
Net dwell time	-2.38	17	.029	-220 (-415 - -25)
Time-to-first fixation	-1.27	17	.22	-152 (-405 - 101)
Reading time per image	-4.30	17	<.001	-4.60 (-6.86 - -2.34)
Sensitivity	3.32	18	.004	.088 (.032 - .14)
Specificity	2.58	18	.019	.076 (.014 - .14)

p = .33, contrast $t(23)$ = -0.42, p = .68. A tendency towards higher total reading time was found for the CEM-groups, $F(2, 25)$ = 3.15, p = .06, $t(24)$ = 1.44, p = .16. On sensitivity, a positive effect in favor of CEM-groups was found, $F(2, 26)$ = 17.2, p < .001, contrast $t(8.09)$ = .003. No effect was found on specificity, $F(2, 26)$ = 2.01, p = .16, contrast $t(24)$ = 1.65, p = .11. Finally, a positive effect in favor of the CEM-groups was found on AUC, $F(2, 26)$ = 16.1, p < .001, contrast $t(24)$ = 5.62, p < .001. In summary, participants in the CEM-groups were not more efficient during the reading process but scored higher on diagnostic performance compared to FFDM.

DISCUSSION

In general, dedicated CEM workstations are configured with a LE-RC hanging protocol (18). Hence, radiologists view the exam by first starting with the LE images, evaluating them as regular FFDM images (5). The RC image is then used to check whether lesions (which were observed on the LE image) enhance or not. Radiologists with experience in the reading of CEM-cases report that lesions are more salient on RC images than on the LE images. By reversing the current hanging protocol, the attention of the radiologist could be immediately drawn to conspicuous areas. Radiologists can thus find lesions up to 31% (6 seconds) faster and can potentially reach higher diagnostic performance if they would evaluate the RC image prior to the LE image. An average decrease in case reading time of 6 seconds may sound small. However, the difference will add up as radiologists may read perhaps tens or even hundreds of cases on a daily basis. Moreover, these differences in reading times were found in radiologists who are experts

in terms of accuracy as well as speed. The differences may even be more pronounced in less experienced radiologists.

For this study, three hypotheses were tested: (1) participants of a (reversed) RC-LE hanging protocol will be more efficient and will score higher on diagnostic performance compared to participants using the (regular) LE-RC hanging protocol; (2) participants will be more efficient in their reading process and will score higher on diagnostic performance of RC images compared to LE images; (3) participants of using any CEM protocol will be more efficient and score higher on diagnostic performance compared to participants using conventional FFDM.

With respect to the order of hanging protocols, no differences were found on eye-movement measurements, nor on diagnostic performance. Nevertheless, a difference in average reading time per case was observed. Consequently, hypothesis (1) was not supported. We assumed that the higher saliency of malignancies on RC images would direct the radiologists` attention towards these areas faster, which would be reflected by more efficient eye movements. However, eye movements in our study proved to be similar in both groups. While eye movements can provide invaluable information about visual search and attention, it does not provide definite answers on how abnormalities are interpreted (19, 20).

The interpretation process after the detection of an abnormality is, to some extent, reflected by the average reading time (13). The difference found in average reading time between the two protocols was mainly caused by the lower average reading time of the LE images, since the reading times of RC images did not differ between the two protocols. In addition, participants of the RC-LE protocol showed identical diagnostic performance compared to the LE-RC protocol. The combination of shorter reading time and equally high diagnostic performance indicates that abnormalities were interpreted more easily by radiologists using the RC-LE protocol compared to the LE-RC protocol.

Second, it was found that participants of CEM-groups fixated less often and shorter on malignancies on RC images compared to LE images. Therefore, hypothesis 2 was supported. Fewer fixations and less time needed for fixating on abnormalities to evaluate a patient case can indicate that the interpretation process of RC images was more efficient (13). Participants did

not fixate earlier on malignant lesions on RC images compared to LE images as average time-to-first fixation did not differ. However, considering the substantial standard deviations, a potential yet small effect between the two image categories may not have been discernible. Furthermore, participants scored higher on diagnostic performance on RC images compared to LE images. However, this observation is less relevant, as RC and LE images are evaluated together in clinical practice.

Third, it was found that the reading processes on LE images were similar to the reading process of FFDM images, confirming previous findings that they are diagnostically equal (5,6). In line with many previous studies, our study also showed that diagnostic performance was superior in CEM groups when compared to conventional FFDM only (1, 2, 5, 21). Therefore, hypothesis 3 was partly supported. LE images are comparable to FFDM images (5), and this similarity presumably caused a similar reading strategy.

The findings of this study may have clinical implications. In the imaging community, there is some concern about the increased reading time of CEM exams, as it consists of double the number of images per patient. This increase might result in substantial increases in workload for radiology staff. Also, Lebron-Zapata et al. (22) showed that CEM might even be considered as a screening tool for women with high risk for developing breast cancer. When high volumes of CEM exams are produced, such as in screening settings, an increase in reading time is not desirable. This investigation shows that these arguments may be less of a concern as long as an RC-LE hanging protocol is used. Moreover, the results indicate that sensitivity could increase with 25% (from 79% to 99%) and specificity with 13% (from 80% to 90%) with similar reading times; when an RC-LE hanging protocol is adopted in a screening setting instead of the current FFDM standard.

To the best of our knowledge, this is the first study on the effects of hanging protocol modification on radiologists` reading processes and diagnostic performance. It is shown that the time to evaluate a case is influenced and is one-third shorter for a particular sequence, while diagnostic performance was not influenced. Modification of hanging protocols could thus impact radiologists` workflow. In most Picture Archiving and Communication Systems (PACS), it is possible to modify hanging protocols, yet it is unknown if radiologists do so and what the effects are. More research on the influence

of modification of hanging protocols in different radiologic examinations is advised.

Perhaps slightly counterintuitive, the findings on the first two research questions indicate that the main value of the RC-image may not lie in the detection of abnormalities, but more in the interpretation of abnormalities. Considering that the radiologists in this research needed fewer fixations and less time to evaluate RC images while scoring higher on diagnostic performance, the RC images can be considered the less complex component of CEM. For training purposes, it is generally recommended to start with less complex material and to gradually increase the complexity as the learner advances (23, 24). Showing CEM cases with the RC image prior to the LE image may help (breast) radiologists in training in learning to read conventional mammograms. Similarly, it is also recommended to start with RC images, followed by LE images when learning to read CEM exams.

This study has some limitations. First, due to the small sample size, some small effects may not have been detectable, such as a potential effect on average time-to-first fixation. In eye-tracking research, groups with a different level of expertise are compared, such as novices and experts (25). Eye-tracking studies like ours, with groups of a comparable expertise level, such as this investigation, are scarce (19, 25, 26). Differences in eye movement measures between groups of a comparable expertise level might be smaller than differences between groups of various expertise levels (26). Moreover, in eye-tracking research, sample sizes are generally small (19, 25), and many efforts have been taken to create a sample size as large as possible.

Second, participants were asked to only look for suspicious lesions, while the task in the clinical workplace is much broader than that, for example, detecting suspicious calcifications. However, previous studies have shown that the added value of CEM for the identification of calcifications is limited (27). Although this limitation may have biased our results, this bias is similar for the three experimental groups.

Another limitation concerns suspicious lesions that do not enhance on contrast-enhanced images, such as mucinous carcinomas or (micro) calcifications (5, 27, 28). Users of CEM should always be aware of other lesions that do not enhance and check extra for these lesions. However,

it could be the case that readers overlook non-enhancing lesions. Cases containing non-enhancing lesions were not used in this experiment. The impact of such lesions on the reading process and diagnostic performance of radiologists using an RC-LE or LE-RC hanging protocol can, therefore, not be deciphered with this experiment. A follow-up study with suspicious yet non-enhancing lesions is warranted.

Finally, in this experiment, participants could not return to a previous image of a patient case as this could blur the findings of the investigation on the order of the hanging protocol. In the clinical workplace, however, radiologists are able to switch between LE and RC images as often as they would like. Therefore, a follow-up study in a more clinical and ecologically valid setting is warranted.

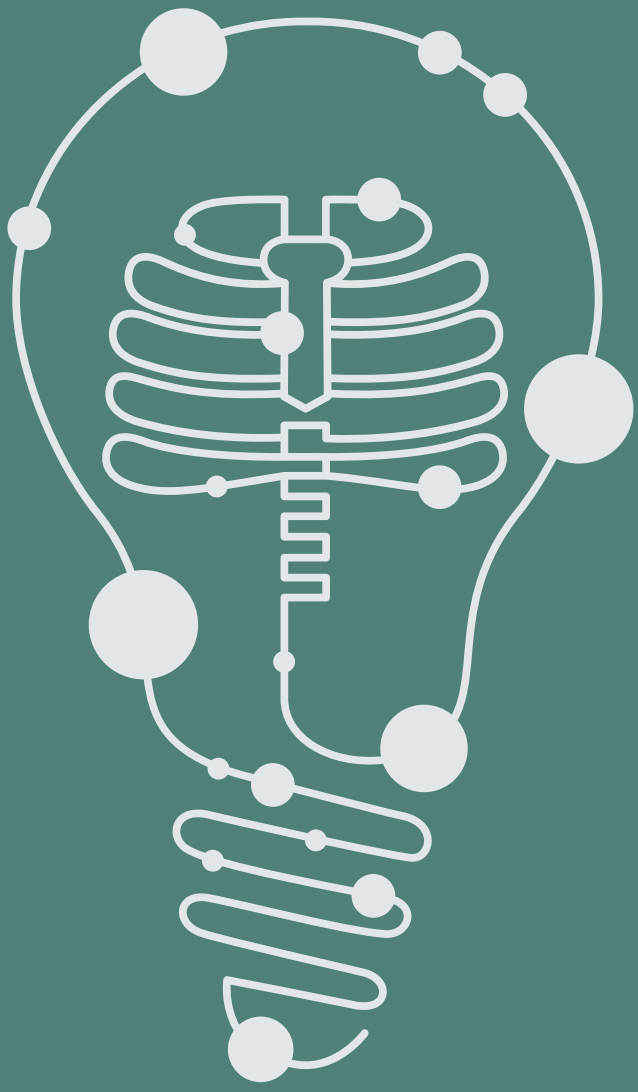
CONCLUSION

Reversal of a CEM hanging protocol from the commonly used LE-RC order to the RC-LE order lowers the case reading time, while diagnostic performance is maintained for breast cancer detection. Furthermore, the reading process is more efficient. Like other studies, we showed that the diagnostic accuracy of CEM is superior to FFDM. Based on our observations, we would recommend using an RC-LE hanging protocol in everyday clinical practice, but also in training.

REFERENCES

1. Patel BK, Lobbes MBI, Lewin J. Contrast Enhanced Spectral Mammography: A Review. Seminars in ultrasound, CT, and MR. 2018;39(1):70-9.
2. Tagliafico AS, Bignotti B, Rossi F, Signori A, Sormani MP, Valdora F, et al. Diagnostic performance of contrast-enhanced spectral mammography: Systematic review and meta-analysis. Breast (Edinburgh, Scotland). 2016;28:13-9.
3. Houben IPL, Van de Voorde P, Jeukens C, Wildberger JE, Kooreman LF, Smidt ML, et al. Contrast-enhanced spectral mammography as work-up tool in patients recalled from breast cancer screening has low risks and might hold clinical benefits. Eur J Radiol. 2017;94:31-7.
4. Lalji UC, Houben IP, Prevos R, Gommers S, van Goethem M, Vanwetswinkel S, et al. Contrast-enhanced spectral mammography in recalls from the Dutch breast cancer screening program: validation of results in a large multireader, multicase study. European radiology. 2016;26(12):4371-9.
5. Lalji UC, Jeukens CR, Houben IPL, Nelemans PJ, van Engen RE, van Wylick E, et al. Evaluation of low-energy contrast-enhanced spectral mammography images by comparing them to full-field digital mammography using EUREF image quality criteria. European radiology. 2015;25(10):2813-20.
6. Francescone MA, Jochelson MS, Dershaw DD, Sung JS, Hughes MC, Zheng J, et al. Low energy mammogram obtained in contrast-enhanced digital mammography (CEDM) is comparable to routine full-field digital mammography (FFDM). Eur J Radiol. 2014;83(8):1350-5.
7. Lobbes M. Comparison Between Breast MRI and Contrast-Enhanced Digital Mammography. 2018. p. 47-56.
8. S. Jochelson M, Pinker K, DD Dershaw, Hughes M, Gibbons GF, Rahbar K, et al. Comparison of screening CEDM and MRI for women at increased risk for breast cancer: A pilot study. Eur J Radiol. 2017 Dec;97:37-43
9. Fallenberg EM, Schmitzberger FF, Amer H, Ingold-Heppner B, Balleyguier C, Diekmann F, et al. Contrast-enhanced spectral mammography vs. mammography and MRI - clinical performance in a multi-reader evaluation. European radiology. 2017;27(7):2752-64.
10. Sorin V, Yagil Y, Yosepovich A, Shalmon A, Gotlieb M, Neiman OS, et al. Contrast-Enhanced Spectral Mammography in Women With Intermediate Breast Cancer Risk and Dense Breasts. AJR Am J Roentgenol. 2018:W1-w8.
11. Nishikawa N, Yanagisawa K, Naoi K, Ohnuma Y, Muramatsu Y. Possibility of Exposure Dose Reduction in Contrast Enhanced Spectral Mammography Using Dual Energy Subtraction Technique : A Phantom Study. 2014; Cham: Springer International Publishing.
12. Markey MK. Physics of mammographic imaging: CRC press; 2012.
13. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, van de Weijer J, et al. Eye Tracking: A Comprehensive Guide to Methods and Measures: Oxford University Press; 2011.
14. Jarodzka H, Holmqvist K, Gruber H. Eye tracking in Educational Science: Theoretical frameworks and research agendas. Journal of Eye Movement Research. 2017;10(1).
15. Kok EM, Jarodzka H. Before your very eyes: the value and limitations of eye tracking in medical education. Med Educ. 2017;51(1):114-22.
16. Bhimani C, Matta D, Roth RG, Liao L, Tinney E, Brill K, et al. Contrast-enhanced Spectral Mammography: Technique, Indications, and Clinical Applications. Academic radiology. 2017;24(1):84-8.
17. Yagil Y, Shalmon A, Rundstein A, Servadio Y, Halshtok O, Gotlieb M, et al. Challenges in contrast-enhanced spectral mammography interpretation: artefacts lexicon. Clin Radiol. 2016;71(5):450-7.
18. Carton A-K, Saab-Puong S, Suminski M. SenoBright Contrast Enhanced Spectral Mammography Technology. General Electric Company, Wauwatosa, WI. 2012:7.
19. Gegenfurtner A, Kok E, van Geel K, de Bruin A, Jarodzka H, Szulewski A, et al. The challenges of studying visual expertise in medical image diagnosis. Med Educ. 2017;51(1):97-104.
20. Kok EM, van Geel K, van Merriënboer JJG, Robben SGF. What We Do and Do Not Know

about Teaching Medical Image Interpretation. Frontiers in Psychology. 2017;8(309).
21. Zhu X, Huang JM, Zhang K, Xia LJ, Feng L, Yang P, et al. Diagnostic Value of Contrast-Enhanced Spectral Mammography for Screening Breast Cancer: Systematic Review and Meta-analysis. Clinical breast cancer. 2018.
22. Lebron-Zapata L, Jochelson MS. Overview of Breast Cancer Screening and Diagnosis. PET clinics. 2018;13(3):301-23.
23. Van Merriënboer JJ, Clark RE, De Croock MB. Blueprints for complex learning: The 4C/ID-model. Educational technology research and development. 2002;50(2):39-61.
24. van Merriënboer JJG, Sweller J. Cognitive load theory and complex learning: Recent developments and future directions. Educational Psychology Review. 2005;17(2):147-77.
25. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. Educational Psychology Review. 2011:1-30.
26. van Geel K, Kok EM, Dijkstra J, Robben SGF, van Merriënboer JJG. Teaching Systematic Viewing to Final-Year Medical Students Improves Systematicity but Not Coverage or Detection of Radiologic Abnormalities. Journal of the American College of Radiology: JACR. 2017;14(2):235-41.
27. Cheung YC, Juan YH, Lin YC, Lo YF, Tsai HP, Ueng SH, et al. Dual-Energy Contrast-Enhanced Spectral Mammography: Enhancement Analysis on BI-RADS 4 Non-Mass Microcalcifications in Screened Women. PloS one. 2016;11(9):e0162740.
28. Houben IP, Vanwetswinkel S, Kalia V, Thywissen T, Nelemans PJ, Heuts EM, et al. Contrast-enhanced spectral mammography in the evaluation of breast suspicious calcifications: diagnostic accuracy and impact on surgical management. Acta Radiologica: Acta Radiol. 2019. 24:284185118822639



Chapter 6

General discussion

6

GENERAL DISCUSSION

In this Ph.D. thesis, it is investigated how learning to evaluate medical images takes place over the range of learners from medical students to radiologists, and its findings can be used to support lifelong learning. As the importance and availability of radiological images in everyday medical practice increases, learning to evaluate medical images has become of great interest to many medical students and non-radiology physicians, residents in radiology, and radiologists. The evaluation of medical images is considered a complex skill requiring many years of practice and training. Different strategies are essential to tailor education to the specific needs for learners of the whole expertise spectrum: For novices, it is relevant to design radiology teaching initiatives as effective and efficient as possible. For intermediates, it is essential to improve understanding of how the evaluation process evolves to provide feedback and to monitor learning. For experts, it is necessary to adapt to new imaging techniques, and the implementation of these new techniques needs to be supported.

The thesis focused on the following research questions. The first and second research questions, 'What are the effects of the prevalence of normal images in a practice phase on third-year medical students' lesion detection and analysis?' and "What are the effects of instructional sequence on third-year medical students' lesion detection and analysis?' were addressed in Chapter 2. The third research question, "Do visual search patterns and lesion detection change after systematic viewing training for final-year medical students?" was studied in Chapter 3. The fourth research question, "How does the longitudinal development of visual search patterns and detection skills of residents in radiology take place?" was addressed in Chapter 4. Finally, in Chapter 5, the fifth research question was investigated: "What are the effects of reversal of the evaluation order of LE and RC mammograms on visual search patterns and malignant lesion detection by breast radiologists?". In this General Discussion, the main findings of these five research questions will be presented. The theoretical contributions of the studies will be discussed, as well as opportunities for future research. The practical implications of the studies will be subsequently presented. Next, some limitations of this thesis are scrutinized. The General Conclusion completes this General Discussion.

MAIN FINDINGS

RQ 1: What are the effects of the prevalence of normal images in a practice phase of medical image evaluation training on third-year medical students` lesion detection and analysis?

A tradeoff between sensitivity and specificity on the post-test was found based on the prevalence of normal and abnormal images during the practice phase: The students practicing with predominantly abnormal images scored higher on the post-test sensitivity while the students practicing with predominantly normal images scored higher on the post-test specificity.

These findings indicate that for novice learners, such as medical students, the proportion of normal and abnormal images should be an important consideration of designing image evaluation training. This proportion affects the learners` sensitivity and specificity, and thus, proportions in line with the prevalence of everyday medical practice are advised to support novice learners optimally.

RQ 2: What are the effects of an inductive and deductive instructional sequence in medical image evaluation training on third-year medical students` lesion detection and analysis?

Inductive sequences are advocated for productive failure, which is known as the act of problem-solving, leading to initial failure, yet eventually resulting in a deeper understanding of the problem. Students' lesion detection and diagnostic performance were similar on the post-test for the deductive (instruction-first) and inductive (practice-first) sequences. Additionally, students of the inductive sequences needed more evaluation time on the post-test. During the practice phase, students' lesion detection and diagnostic performance were lower for the inductive sequences, and unexpectedly, evaluation times were shorter in the inductive sequences. Moreover, students needed more evaluation time for erroneously interpreted cases during the practice phase in both instructional sequences. Thus, productive failure probably occurred in inductive as well as deductive conditions. Additionally, no evidence was found that inductive sequences lead to a deeper understanding as both sequences led to similar post-test diagnostic performance scores.

When designing image evaluation training, expert instruction prior to a practice phase (a deductive sequence) is advised, even for novice learners in radiology. A deductive sequence for novices may seem contradictory to findings of previous investigations, where it was found that novices particularly benefit from inductive instructional sequences. However, while these third-year medical students can be considered novices in image evaluation, they already built up some knowledge basis on anatomy, physiology, and pathology and cannot be considered completely unscathed laypeople. Thus, deductive sequences for novices and more advanced learners in radiology are advised.

RQ 3: Do visual search patterns change, and does the detection of abnormalities increase after systematic viewing training of final-year medical students when evaluating chest radiographs?

The effects of a systematic viewing training on coverage, systematicity, the detection of lesions were compared to a nonsystematic training for the evaluation of chest radiographs by final-year medical students in Chapter 3. It was found that the students' coverage similarly increased in both conditions, yet only the students of the systematic viewing group became more systematic after their respective training. Moreover, the detection of abnormalities increased post-training in both conditions similarly. Overall, this study's findings do not support the claim that systematic viewing training intrinsically augments the image evaluation skills of novices in radiology.

To support novice learners' image evaluation, teachers should emphasize the visual information of medical images, such as anatomy and potential abnormalities, instead of viewing strategies. Systematic viewing training is generally advocated because of the assumption that being more systematic leads to more complete evaluations and increased detection of abnormalities. This assumption does not hold as the post-training coverage and detection of abnormalities increased similarly in the systematic viewing and the nonsystematic viewing condition. The training videos of both conditions covered the same visual information, consisting of the anatomy and potential abnormalities visualized on chest radiographs. Thus, the similar increase in coverage and detection of abnormalities of both conditions most likely resulted from the provided visual information and not by providing a (systematic) viewing strategy per se.

RQ 4: How does the longitudinal development of visual search patterns and lesion detection skills of first-year residents in radiology take place when evaluating chest radiographs?

The longitudinal development of evaluation time, visual search patterns, and lesion detection of first-year residents in radiology on chest radiographs were investigated in Chapter 4. Evaluation times were halved in only four months, accompanied by more efficient visual search patterns: Over time, fewer fixations, a gradual decrease of average fixation durations, and increases in mean saccade lengths were found. Finally, the proportion of abnormal dwell time increased with the largest changes occurring during the first four months of training. These changes in visual search patterns support the most popular theories on visual expertise development (1, 2).

The findings of this study also indicate that the residents already adapted their visual search to image characteristics. When abnormal images were compared to normal images, longer evaluation times, lower numbers of fixations, and longer mean saccade lengths were found, indicative of an adaptation of visual search to image characteristics. Moreover, these differences in visual search patterns on abnormal compared to normal images further indicate that image characteristics should be taken into account when visual search patterns are investigated.

Finally, lesion detection remained constant over the year, yet evaluation times significantly decreased. Therefore, the longitudinal effect of training on lesion detection is probably indirect via decreasing evaluation times. Additionally, specificity was lower on abnormal cases compared to normal cases; The participants thus made more false-positive errors on abnormal cases.

Residents primarily engage in workplace learning (3) and learn through feedback on their own image evaluations. For intermediate learners such as residents, feedback on image evaluations as well as monitoring of development, are thus central to their learning experiences (3). It was shown that eye-tracking methodology could provide new and in-depth information about the development of image evaluation skills of residents. Eye-tracking can thus uncover a part of the learning process previously not seen. The findings of this longitudinal study can have added value for the monitoring of learning processes. For example, evaluation time,

the number of fixations, and the proportion abnormal dwell time showed the largest changes during the first four months of training. When these parameters are still pronouncedly changing in evaluating a specific medical image, such as chest radiographs, the resident is probably still developing visual search. In such cases, it might be advisable to prolong the exposure to this medical image type.

RQ 5: What are the effects of reversal of the evaluation order of plain (LE) and contrast-enhanced (RC) mammograms on visual search patterns and the detection of malignant breast lesions by breast radiologists?

The evaluation order of a contrast-enhanced (RC) mammogram followed by a low-energy (LE) mammogram (RC-LE order) was compared to the traditional LE-RC order. The effects of evaluation order on evaluation time, visual search patterns, and the detection of malignant breast lesions were investigated to improve understanding of how experts could implement new imaging techniques into everyday clinical practice. Moreover, the RC-LE and LE-RC orders were compared to a LE-only (similar to plain mammograms) group to investigate the addition of contrast-enhanced mammograms on evaluation time, visual search patterns, and lesion detection.

Evaluation time was 33% lower for the RC-LE order compared to the LE-RC order, while lesion detection was similar in both conditions. The eye-tracking parameters did not significantly differ between the RC-LE and LE-RC conditions. Thus, the reversal of the evaluation order may particularly impact the analysis instead of the detection of any anomalous area. Finally, evaluation time did not significantly differ between the RC-LE and LE-RC conditions compared to plain mammograms only. However, the detection of malignant breast lesions was superior for the combined RC-LE and LE-RC conditions.

Radiologists, while they are the acknowledged experts in their field, still need to adapt continuously: They need to keep up with the constant progression of the radiology field with the constant introduction of new imaging techniques. This study was meant to showcase of how radiologists could be supported to adapt to new imaging techniques by investigating the evaluation process. The eye-tracking methodology proved beneficial to uncover a part of the previously covert evaluation process. Therefore,

investigations with eye-tracking methodology are advised to tailor the implementation of new imaging techniques into the radiology field.

THEORETICAL CONTRIBUTIONS

Based on the main findings of the separate studies, three theoretical contributions were identified in learning to evaluate medical images for the range of learners from medical students to radiologists. These contributions will now be discussed, and future research opportunities will be delineated. First, the lessons learned on teaching medical image evaluation will be presented, followed by a treatise on the impact of normal images on learning and future practice. Finally, the added value of eye-tracking methodology to provide insights on evaluation processes will be considered.

Lessons learned on teaching radiology

When building expertise in radiology, one has to learn how to search for and how to analyze abnormalities (4, 5). It is generally assumed that the processes of searching and analyzing abnormalities run parallel, and not serial, during the evaluation of radiological images. To put this to extremes, it is impossible to analyze something that one cannot find and to search for something without knowing how it looks. Therefore, both processes most likely influence each other (4-6), even at early phases of expertise development. Both of these aspects have been researched and lead to different considerations for teaching radiology or medical image evaluation in general.

Considering the question of how to search, it was found that systematic viewing training before the actual evaluation of images was not beneficial for the detection of abnormalities, neither in the study presented in Chapter 3 nor in other experimental studies with balanced experts' explanation (7, 8). Should a teacher thus completely abandon the topic of how to search in radiology education? Perhaps not, students are generally interested in developing a systematic approach to evaluate radiological images as it can provide guidance (7, 9). However, considering the intricate interplay between searching and analyzing images, an interesting avenue of future research would be to optimize this guidance during evaluations, instead of before evaluating radiological images. Nowadays, most Picture Archiving and Communication Systems (PACS) provide opportunities for standardized or structured reporting (10, 11). A structured radiology report could contain all the relevant anatomical structures of an image. The novice should

report on all of these structures of the structured report while evaluating the image. Structured reporting could thus safeguard complete evaluations and provide essential guidance for novices during their evaluation process (11). Finally, since searching and analyzing are interconnected, learning how to search should always be combined with learning how to analyze abnormalities.

Considering how to analyze abnormalities, what should be the subject and design of image evaluation training? Regarding the subject of image evaluation training for novices, a greater emphasis on normal images is advised, further delineated in the second theoretical contribution. Regarding the design of image evaluation training, neither the current literature on teaching medical image evaluation nor the studies reported in this thesis can provide clear-cut answers on how to design the most effective and efficient learning experience (12): In previous studies on teaching medical image evaluation, generally new (e-learning) courses were compared to the old ones. When courses are redesigned, many aspects change concurrently, making it impossible to attribute effects to distinctive aspects. To avoid these flaws in experimental educational research, similar to experimental studies in other fields, potential confounders should be kept constant to attribute effects to specific factors, such as the effects of inductive and deductive instructional sequences. While some investigations used an experimental approach, such as the studies presented in Chapters 2 and 3, this body of evidence is currently not sufficient to provide the necessary answers on how to design image evaluation training. Further research is warranted and should predominantly have an experimental approach with potential confounders kept constant.

Additionally, and as a final remark on researching radiology education in general, it is advised that such research is rooted in educational and learning theories (12-14). Many educational interventions and instructional designs have already been researched in other educational fields. Their findings may apply to radiology education. Radiology education can particularly benefit from educational research in other domains with a large visual component, such as pathology (15), reading electrocardiograms (16), dermatology, and even domains outside the medical field such as air traffic control (17). While further research is generally warranted, in the case of radiology education, a literature search on previous -- educational or psychological -- research is warranted.

The impact of normal images on learning and future practice

How does one decide that a radiological image does not contain any abnormalities and should thus be considered normal? Many non-radiology physicians will have to differentiate normal from abnormal findings every day, which is a crucial component of image evaluation (18). An erroneous differentiation may have serious consequences (19): False-positive findings will lead to unnecessary additional procedures, whereas false-negatives can lead to delays in diagnosis. While both situations are unfavorable, it depends on the context which scenario is the least harmful. Modification of this differentiation process may help to shift the balance towards the most favorable situation. Theories, such as signal detection theory (20), can provide insights into how differentiation processes take place and how they could be modified.

According to signal detection theory (SDT), radiological images will always contain "noise" (21, 22). This noise may come from the image itself, e.g., imaging artifacts, but also from patients (23). Examples of patient noise are large vessels on a chest radiograph mimicking chest nodules or dense breast tissue mimicking malignant lesions on a mammogram. The appearance of benign or normal findings can be almost identical to the appearance of abnormal, malignant findings (24). Therefore every evaluation, even by senior radiologists, will inevitably have some level of uncertainty (22). The two key concepts in SDT that influence the differentiation of normal from abnormal findings are the evaluators' differentiating ability and the evaluators' susceptibility to call a finding normal, also known as the criterion (20-22). Both the evaluators' ability and the criterion can be modified to influence the differentiation process (22).

Modifying future non-radiology physicians' ability may not be a realistic goal for medical curricula (25). The ability to differentiate normal from abnormal is highly related to the evaluators' expertise level, which can be considered low for non-radiology physicians (such as junior doctors) generally working on a ward or emergency department (26). Therefore, modifying the ability to differentiate is nothing more than expertise development and will require extensive practice and teaching with thousands of cases (27). As medical curricula are already brimmed, there will not be sufficient time and resources for such learning activities (18, 25).

Modification of the criterion of non-radiology physicians may be more opportune. Criterions can be influenced by the proportion of normal and abnormal images in a radiology teaching initiative: In Chapter 2, a criterion shift in novices was found, based on different proportions of normal and abnormal cases in a practice phase. This practice phase consisted of only 20 chest radiographs. Criterion shifts based on prevalence have also been established in other image evaluation training (21, 28). Criterions may even change per image; significantly lower specificity scores on abnormal images compared to normal images were found in the longitudinal study reported in Chapter 4. The residents participating in this study may have shifted towards a more tolerant criterion when identifying one abnormality and could have searched for additional abnormalities as they already may have learned that many cardiopulmonary diseases have multiple radiological manifestations (5, 29-31).

However, the lower specificity scores -- and thus higher false-positive rates -- on abnormal images of the longitudinal study may seem contradictory to the literature on Satisfaction of Search (SOS) errors (32, 33). SOS is defined as the premature closure of visual search due to the identification of one abnormality, which should have resulted in lower false-positive findings on abnormal images instead of higher false-positive findings in the longitudinal study. However, SOS is generally thought to interfere with the visual search for findings unrelated to the diagnosis, such as the miss of a lung nodule when diagnosing pneumonia in a patient experiencing shortness of breath and fever (33). The higher false-positive findings on abnormal cases of the longitudinal study may be explained by a more tolerant criterion caused by a search for related abnormalities, such as the search for rib fractures after the identification of pneumothorax. Nonetheless, it is challenging to distinguish visual search for related and unrelated findings with eye-tracking methodology alone. Further research on the analysis of abnormalities with additional outcome measures is advised.

Teachers can use the effect of prevalence to optimize criterions for future practice. Additionally, it is estimated that criterion shifts based on the prevalence effect only subside after approximately 50 trials in luggage screening, another visual domain where the prevalence of search targets is generally low (31, 34). Criterion shifts of novices based on the proportion of normal and abnormal cases in training settings different from the prevalence of medical practice may thus not be resolved easily. Novices generally

acquire a fairly accurate sense of the prevalence of training cases (21). A proportion of normal and abnormal images in training settings aligned with the prevalence of abnormalities in medical practice is advised.

Progress by measuring the evaluation process

More insights into how the evaluation process takes place have added value for the whole spectrum from novices to senior radiologists. Investigating the evaluation process can reveal how novices change their visual search after image evaluation training, how intermediates longitudinally develop their evaluation skills and, inform senior radiologists how to use a new imaging technology in clinical practice. Evaluation processes start with a visual search for abnormalities, and eye-tracking methodology can be used to investigate a substantial share of this process (35-37).

One could also ask an evaluator where he/she is looking when evaluating an image. However, in contrast to eye-tracking methodology, this may not be considered objective information since what people say and what people do is generally different when dealing with cognitive processes, such as a visual search for abnormalities (38). Indeed, reporting on own eye movements has been generally found to be unreliable (39-41). Eye-tracking provides more objective and accurate measures where one is looking and can thus uncover the covert cognitive process of visual search for abnormalities (42).

Additionally, one could also advocate using a stopwatch to study evaluation processes. Indeed, pronounced changes in evaluation times were found in the longitudinal study of Chapter 4 and the study on the reversal of the evaluation order of Chapter 5. However, additional and objective measures such as eye-tracking parameters can provide a much more detailed picture of the evaluation process and explain why changes occur. The longitudinal increase in the proportion of abnormal dwell time, combined with the lower evaluation times in the study reported in Chapter 4, is indicative that visual search becomes more efficient with increasing expertise. Likewise, the similar eye-tracking parameters of the RC-LE and LE-RC order of Chapter 5 indicate that the reversal of the evaluation order did not lead to incomplete evaluations. In both cases, eye-tracking measures explained why changes in the evaluation process occurred.

Investigating the process itself can provide new and invaluable information for learning and integrating new techniques into the everyday workflow, yet the outcome of the process should not be left out of the equation. Process measures and outcome measures are both essential to discern whether a visual search is efficient and effective. For example, there may be four students evaluating a chest radiograph. This chest radiograph has a tumor at the apex of the right upper lung lobe and a prominent gastric bubble under the left diaphragm, which is considered a normal finding. The first student correctly identified the tumor, and his/her visual search is thus effective. The process measures, such as eye-tracking measures, also indicated efficient search as this student did not have sustained attention for the normal features, such as the gastric bubble. The second student identified the tumor too, and the search was therefore also effective. However, the process measures indicated that while the search was effective, it could also be considered inefficient; this student particularly focused on the salient gastric bubble. The third student already showed efficient search, without sustained attention on the gastric bubble, yet this efficient search did not lead to improved outcome already as he/she failed to identify the tumor. Finally, the fourth student was found neither efficient nor effective as a strenuous search did not lead to the identification of the tumor. While two students are effective, and two are efficient in this example, all four students should receive completely different feedback on their evaluations to thrive their learning. To provide the most fertile feedback to evaluators, both the process and outcome measures are equally essential. Thus, when processes are investigated, the outcome should always be taken into account.

Furthermore, the particular medical image type should always be taken into account as well when interpreting eye-tracking parameters (36). Aggregation of eye-tracking data from numerous studies in radiology (37), or beyond the boundaries of medicine (43) may suggest that the development of visual expertise always follows the same pattern. This assumption most likely does not hold (36). It has been found that eye-tracking parameters differ between two-dimensional and volumetric image types (37, 40), on different disease patterns for the same image type (36, 44) and even on the same image when different instructions are provided (45). Therefore, visual expertise may also be understood as the ability to adapt the visual search to image characteristics (36), and it is therefore essential to take the image characteristics into account when studying visual search patterns with eye-tracking methodology.

PRACTICAL IMPLICATIONS

Radiology teachers are encouraged to pay particular attention to the prevalence of normal and abnormal images in image evaluation training. In image evaluation training, there is currently an emphasis on abnormal findings and diagnosing radiological images (18, 25). Diagnosing medical images is a task that non-radiology physicians will most likely not succeed in mastering (25, 26, 46). Medical students may feel overwhelmed by all potential abnormal findings in image evaluation training (9, 25), and diagnosing images could thus be considered too complex for first radiology learning experiences. Moreover, normal radiological images are generally considered less complex to evaluate compared to abnormal images (47), and should thus be the starting point for image evaluation training. Furthermore, radiological images are predominantly normal in everyday clinical practice (48, 49), and centering radiology education on the variety of normal findings may also improve the preparation of medical students for future practice. After the completion of medical school, these new junior doctors will eventually encounter abnormal cases and expand their knowledge basis on image evaluation more gradually through workplace learning.

Process measures, such as eye-tracking parameters, can provide new and in-depth information about the evaluation process to intermediate and advanced learners in radiology, such as residents and radiologists. Process measures could enrich the current feedback to residents. This feedback could enhance learning in residency training, ideally when combined with large radiological image banks. For example, a resident may find it difficult to diagnose subtle pneumothoraxes. The eye-tracking parameters are indicative of faulty searches, the resident may discuss the faulty search strategy with a supervisor to improve the search strategy, and the resident could subsequently practice with pneumothorax cases from the image bank. A new evaluation session with eye-tracking could confirm any improvements. Anecdotally, some participants of the longitudinal study reported in Chapter 4 were eager to receive feedback based on their eye-tracking parameters and lesion detection and wanted to know whether they were missing particular lesions.

Eye-tracking investigations produce considerable amounts of data (50) that can be strenuous for lay-people to interpret. In order to use eye-tracking measures as feedback, they thus need to be processed for laypeople into

visualizations to provide rich and in-depth feedback to learners. Such visualizations could be learning curves and medical images, such as a chest radiograph, with the learners` fixations superposed to indicate erroneously omitted areas in visual search. Such visualizations can require substantial resources. Artificial intelligence could help to transform the abundant eye-tracking information into such visualizations.

One step further, artificial intelligence could eventually result in augmented radiology training. Augmented radiology training is analogous to future augmented radiology practice, where computers will support humans to produce the most fruitful results (51, 52). Augmented radiology training could consist of an automated feedback system based on the eye-tracking and outcome measures, and it could also help in the selection of cases from an image bank to make the training adaptive to individual learner's needs. Such a system could foster deliberate practice, one of the hallmarks of expertise development (53).

Furthermore, the results of the CEM-study in Chapter 5 indicate that process measures can provide in-depth information on how a new imaging technique can be implemented in clinical practice optimally. CEM is currently used as a secondary imaging technique for women recalled from breast cancer screening. However, it is debated to use CEM as a primary imaging technique in breast cancer screening (54). Our CEM-study results add to this debate that CEM could be used as a quick and reliable screening technique if the contrast-enhanced images are evaluated first (RC-LE order).

LIMITATIONS

Some limitations of this Ph.D. thesis should be mentioned. First, all the radiological images used in the experiments were two-dimensional. Visual search and analysis may differ between volumetric and two-dimensional images (5, 51, 55, 56). Volumetric images contain more visual information than two-dimensional images, as volumetric images consist of hundreds or even thousands of slices. Therefore, the visual search component could be larger in the evaluation of volumetric images compared to two-dimensional images (57). Therefore visual search training on volumetric images may be more beneficial for novices compared to the results on two-dimensional images (58). However, chest radiographs and plain mammograms represent a substantial proportion of the radiology departments' workload (59) and are both considered difficult to master (60, 61). Both image types are

thus excellent materials to study expertise development. Nonetheless, any extrapolation of the findings on the two-dimensional experimental material of this Ph.D. thesis to volumetric images should be done with care (36, 37, 55).

Second, eye-tracking research could particularly benefit from a triangulation with other methodologies, such as verbal data, to get an even richer picture of the evaluation process. Eye-tracking data provides invaluable information about the subconscious, covert aspect of the evaluation process, for example, where one has looked (39). Verbal data, such as think-aloud protocols or retrospective reports (62), can supplement this data and provide more insights on why someone has looked at a particular location and how a particular area was analyzed (42, 63). For an even more detailed description of the evaluation process, a triangulation of eye-tracking with other methodologies is advised.

Third, the studies on the effects of systematic-viewing training and the effects of the proportion of abnormal and normal images and instructional sequence only administered immediate post-tests. The absence of a delayed post-test on the investigation of systematic viewing training may be of lesser concern as systematic viewing training did not prove beneficial for lesion detection on an immediate post-test. On the other hand, inductive instructional designs are advocated for the retention of knowledge through productive failure (64). Productive failure could also have occurred in the deductive conditions. Nonetheless, it remains possible that particularly inductive sequences have productive effects on retention of knowledge of image evaluation training.

CONCLUSIONS

In conclusion, different educational strategies are essential to support learners from the whole expertise spectrum in radiology. For novices, image evaluation training should focus less on teaching visual search strategies and more on normal findings instead of abnormal findings. For intermediates, studying the evaluation process with eye-tracking methodology can provide new and fine-grained information for feedback and monitoring. For experts, eye-tracking methodology can provide insights into how to add new imaging techniques to their current clinical practice.

REFERENCES

1. Kundel HL, Nodine CF. A Visual Concept Shapes Image Perception. Radiology. 1983;146(2):363-8.
2. Reingold EM, Sheridan H. Èye movements and visual expertise in chess and medicine. In: Leversedge SP, Gilchrist ID, Everling S, editors. Oxford Handbook on Eye Movements. Oxford: Oxford University Press; 2011. p. 528-50.
3. Kilminster S, Cottrell D, Grant J, Jolly B. AMEE Guide N° 27: Effective educational and clinical supervision. Medical teacher. 2007;29:2-19.
4. Samei E, Krupinski EA. The Handbook of Medical Image Perception and Techniques: Cambridge University Press; 2010.
5. Gijp A, Schaaf MF, Schaaf IC, Huige JCBM, Ravesloot CJ, Schaik JPJ, et al. Interpretation of radiological images: towards a framework of knowledge and skills. Advances in Health Sciences Education. 2014:1-16.
6. Tourassi G, Voisin S, Paquit V, Krupinski E. Investigating the link between radiologists' gaze, diagnostic decision, and image content. J Am Med Inform Assoc. 2013.
7. Kok E, Jarodzka H, Bruin AB, Bin Amir H, Robben SGF, Merrienboer JG. Systematic viewing in radiology: seeing more, missing less? Adv Health Sci Educ Theory Pract. 2015.
8. Varvaroussis DP, Kalafati M, Pliatsika P, Castrén M, Lott C, Xanthos T. Comparison of two teaching methods for cardiac arrhythmia interpretation among nursing students. Resuscitation. 2014;85(2):260-5.
9. Subramaniam RM, Beckley V, Chan M, Chou T, Scally P. Radiology curriculum topics for medical students: students' perspectives. Academic radiology. 2006;13(7):880-4.
10. Rankin RN. Technologies for Teaching: exploring the use of PACS, databases and Teaching files. In: Chhem RK, Hibbert KM, van Deven T, editors. Radiology Education The scholarship of Teaching and learning. Berlin Heidelberg: Springer-Verlag; 2009.
11. Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. Insights into Imaging. 2020;11(1):10.
12. Kok EM, van Geel K, van Merriënboer JJG, Robben SGF. What We Do and Do Not Know about Teaching Medical Image Interpretation. Frontiers in Psychology. 2017;8(309).
13. Norman G. Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. Advances in Health Sciences Education. 2014;19(1):1-5.
14. van Merriënboer JJG, de Bruin ABH. Research Paradigms and Perspectives on Learning. In: Spector JM, Merrill MD, Elen J, Bishop MJ, editors. Handbook of Research on Educational Communications and Technology. New York, NY: Springer New York; 2014. p. 21-9.
15. Jaarsma T, Jarodzka H, Nap M, van Merrienboer JJG, Boshuizen HPA. Expertise under the microscope: processing histopathological slides. Medical Education. 2014;48(3):292-300.
16. Sibbald M, Merrienboer JG, Bruin AB. Is dat your final answer? : How doctors should check decisions: Maastricht University 2013.
17. Meeuwen LV, editor Visual problem solving and self-regulation in training air traffic control2013.
18. Zwaan L, Kok E, van der Gijp A. Radiology education: a radiology curriculum for all medical students? Diagnosis. 2017;4.
19. Berlin L. Radiologic errors and malpractice: a blurry distinction. AJR Am J Roentgenol. 2007;189(3):517-22.
20. Green DM, Swets JA. Signal detection theory and psychophysics1966.
21. Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. Journal of experimental psychology General. 2007;136(4):623-38.
22. Boutis K, Pecaric M, Seeto B, Pusic M. Using signal detection theory to model changes in serial learning of radiological image interpretation. Advances in health sciences education. 2010;15(5):647-58.
23. Krupinski EA, Berger WG, Dallas WJ, Roehrig H. Searching for nodules: What features attract attention and influence detection? Academic radiology. 2003;10(8):861-8.
24. Keats TE, Anderson MW. Atlas of normal roentgen variants that may simulate disease. 2013.
25. Naeger DM, Webb EM, Zimmerman L, Elicker BM. Strategies for incorporating radiology

into early medical school curricula. Journal of the American College of Radiology : JACR. 2014;11(1):74-9.
26. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department: is the radiologist really necessary? Postgraduate medical journal. 2003;79(930):214-7.
27. Ericsson KA, Krampe RT, Teschromer C. The role of deliberate practice in the acquisition of expert performance. Psychological Review. 1993;100(3):363-406.
28. Pusic M, Pecaric M, Boutis K. How much practice is enough? Using learning curves to assess the deliberate practice of radiograph interpretation. Acad Med. 2011;86(6):731-6.
29. Adam A, Dixon AK, Gillard JH, Schaefer-Prokop C, Allison DJ, Grainger RG. Grainger & Allison's diagnostic radiology : a textbook of medical imaging. 2015.
30. Norman GR, Coblentz CL, Brooks LR, Babcook CJ. Expertise In Visual Diagnosis - A Review Of The Literature. Academic Medicine. 1992;67(10):S78-S83.
31. Ishibashi K, Kita S, Wolfe JM. The effects of local prevalence and explicit expectations on search termination times. Atten Percept Psychophys. 2012;74(1):115-23.
32. Berbaum KS, Krupinski EA, Schartz KM. Satisfaction of search in chest radiography 2015. Academic radiology. 2015;11:1457.
33. Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive Error in Radiology. American Journal of Roentgenology. 2016;208(4):739-49.
34. Wolfe JM, Van Wert MJ. Varying target prevalence reveals two dissociable decision criteria in visual search. Current biology : CB. 2010;20(2):121-4.
35. Kok EM, Jarodzka H. Before your very eyes: the value and limitations of eye tracking in medical education. Med Educ. 2017;51(1):114-22.
36. Bertram R, Kaakinen J, Bensch F, Helle L, Lantto E, Niemi P, et al. Eye Movements of Radiologists Reflect Expertise in CT Study Interpretation: A Potential Tool to Measure Resident Development. Radiology. 2016;281(3):805-15.
37. van der Gijp A, Ravesloot CJ, Jarodzka H, van der Schaaf MF, van der Schaaf IC, van Schaik JPJ, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education. 2017;22(3):765-87.
38. van Merrienboer JJG. What people say # what people do. Perspectives on medical education. 2015;4(1):47-8.
39. Clarke AD, Mahon A, Irvine A, Hunt AR. People are unable to recognize or report on their own eye movements. Quarterly journal of experimental psychology (2006). 2017;70(11):2251-70.
40. Aizenman A, Drew T, Ehinger KA, Georgian-Smith D, Wolfe JM. Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study. Journal of medical imaging (Bellingham, Wash). 2017;4(4):045501-.
41. Võ ML, Aizenman AM, Wolfe JM. You think you know where you looked? You better look again. Journal of experimental psychology Human perception and performance. 2016;42(10):1477-81.
42. Gegenfurtner A, Kok E, van Geel K, de Bruin A, Jarodzka H, Szulewski A, et al. The challenges of studying visual expertise in medical image diagnosis. Med Educ. 2017;51(1):97-104.
43. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. Educational Psychology Review. 2011:1-30.
44. Kok EM, de Bruin ABH, Robben SGF, van Merriënboer JJG. Looking in the Same Manner but Seeing it Differently: Bottom-up and Expertise Effects in Radiology. Applied Cognitive Psychology. 2012;26(6):854-62.
45. Yarbus AL. Eye Movements and Vision. New York: Plenum Press; 1967.
46. Christiansen JM, Gerke O, Karstoft J, Andersen PE. Poor interpretation of chest X-rays by junior doctors. Danish medical journal. 2014;61(7):A4875.
47. Boutis K, Cano S, Pecaric M, Welch-Horan TB, Lampl B, Ruzal-Shapiro C, et al. Interpretation difficulty of normal versus abnormal radiographs using a pediatric example. Canadian medical education journal. 2016;7(1):e68-77.

48. Ng JJ, Taylor DM. Routine chest radiography in uncomplicated suspected acute coronary syndrome rarely yields significant pathology. Emergency medicine journal : EMJ. 2008;25(12):807-10.
49. Verma V, Vasudevan V, Jinnur P, Nallagatla S, Majumdar A, Arjomand F, et al. The utility of routine admission chest X-ray films on patient care. Eur J Intern Med. 2011;22(3):286-8.
50. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Weijer J. Eye Tracking: A Comprehensive Guide to Methods and Measures: Oxford University Press; 2011.
51. Waite S, Farooq Z, Grigorian A, Sistrom C, Kolla S, Mancuso A, et al. A Review of Perceptual Expertise in Radiology-How it develops, How we can test it, and Why humans still matter in the era of Artificial Intelligence. Academic radiology. 2020;27:26-38.
52. Tajmir SH, Alkasab TK. Toward Augmented Radiologists: Changes in Radiology Education in the Era of Machine Learning and Artificial Intelligence. Academic radiology. 2018;25(6):747-50.
53. Ericsson KA, Charness N, Feltovich P, Hoffman RR. The Cambridge Handbook of Expertise And Expert Performance: Cambridge University Press; 2006.
54. Sung JS, Lebron L, Keating D, D'Alessio D, Comstock CE, Lee CH, et al. Performance of Dual-Energy Contrast-enhanced Digital Mammography for Screening Women at Increased Risk of Breast Cancer. Radiology. 2019;293(1):81-8.
55. Venjakob A, Mello-Thoms C. Review of prospects and challenges of eye tracking in volumetric imaging. JMIOBU. 2015;3:011002.
56. Rubin GD, Roos JE, Tall M, Harrawood B, Bag S, Ly DL, et al. Characterizing Search, Recognition, and Decision in the Detection of Lung Nodules on CT Scans: Elucidation with Eye Tracking. Radiology. 2015;274(1):276-86.
57. Drew T, Vo ML-H, Olwal A, Jacobson F, Seltzer SE, Wolfe JM. Scanners and drillers: Characterizing expert visual search through volumetric images. Journal of Vision. 2013;13(10).
58. van der Gijp A, Vincken KL, Boscardin C, Webb EM, Ten Cate OTJ, Naeger DM. The Effect of Teaching Search Strategies on Perceptual Performance. Academic radiology. 2017;24(6):762-7.
59. Lu Y, Zhao S, Chu PW, Arenson RL. An update survey of academic radiologists' clinical productivity. Journal of the American College of Radiology : JACR. 2008;5(7):817-26.
60. Delrue L, Gosselin R, Ilsen B, Landeghem A, De Mey J, duyck p. Difficulties in the Interpretation of Chest Radiography. 2011. p. 27-49.
61. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. 2009;253(3):632-40.
62. Ericsson KA, Simon HA. Protocol Analysis: Verbal Reports As Data: Mit Press; 1993.
63. Jarodzka H, Holmqvist K, Gruber H. Eye tracking in Educational Science: Theoretical frameworks and research agendas. Journal of Eye Movement Research. 2017;10(1).
64. Kapur M. Productive Failure. Cognition and Instruction. 2008;26(3):379-425.

Summary

S

Chapter 1: General introduction

The role of medical images is ever more increasing in everyday medical practice. Furthermore, a growing number of (non-radiology) physicians evaluate images themselves nowadays, and new imaging techniques regularly become available. The evaluation of medical images is considered a complex skill that takes extensive practice and training to master. Novices, residents, and experts have different learning experiences throughout their development of image evaluation skills. They, therefore, need different strategies to support their learning experience. This Ph.D. thesis aims to support lifelong learning by investigating how learning to evaluate medical images takes place for a range of learners of medical students to senior radiologists.

Novices, such as medical students, generally have their first encounters with evaluating medical images in training settings. To optimally support novices, it is relevant to investigate how to provide effective and efficient image evaluation training. The effects of the prevalence of normal and abnormal images and educational (instructional) design on the detection and the analysis of lesions by novices on chest radiographs are investigated in Chapter 2. The effects of a systematic-viewing training on visual search patterns and lesion detection of novices on chest radiographs are investigated in Chapter 3.

Intermediates, such as residents in radiology, engage in workplace learning and thus primarily learn through feedback on their image evaluations. More insights into how the evaluation process takes place could provide residents with feedback and help them monitor their learning. The development of visual search patterns and lesion detection of first-year residents on chest radiographs in radiology is investigated in Chapter 4.

Finally, while radiologists are the acknowledged experts in medical image interpretation, they will still need to continuously develop their image evaluation skills as new imaging techniques frequently become available. Radiologists may wonder how to implement these new techniques into everyday medical practice optimally. The effects of the evaluation order of traditional (plain) and new (contrast-enhanced) mammograms on visual search patterns and malignant lesion detection by breast radiologists are investigated in Chapter 5.

Chapter 2: The effects of prevalence and educational design on lesion detection and analysis by third-year medical students

Image evaluation training predominantly focuses on abnormal findings, while images in medical practice are predominantly normal. This mismatch between training and medical practice may lead to a wrong impression of the prevalence of diseases to students. Moreover, in image evaluation training, novices generally receive expert instruction prior to practice (deductive sequences) and will only apply a told solution. However, for novices, practice prior to instruction is advised (inductive sequences). Such sequences should invoke productive failure; students will need to figure out solutions by themselves and will initially fail yet become fully immersed in problem-solving, eventually leading to a deeper understanding. In a 2x2 between-subjects design, the impact of prevalence (70% normal versus 30% normal cases in a practice phase) and instructional sequence (inductive versus deductive sequences) on lesion detection and lesion analysis of chest radiographs by third-year medical students (n=103) was investigated. A sensitivity-specificity tradeoff was found based on the practice phases' prevalence: students practicing with predominantly abnormal images found more lesions (higher sensitivity). Students practicing with predominantly normal images were more likely to correctly call normal images 'normal' on the posttest (higher specificity). Furthermore, the students of the inductive conditions had similar posttest lesion detection and lesion analysis compared to the students of the deductive conditions. Unexpectedly, students of the inductive conditions took less time per case during the practice phase and could not have explored the cases in enough depth. Furthermore, productive failure was probably also invoked in the deductive conditions and may not be confined to inductive conditions. Overall, for novices' image evaluation training, the proportion of abnormal and normal cases should be an important consideration, and deductive instructional sequences are advised.

Chapter 3: the effects of a systematic-viewing training on visual search patterns and lesion detection of final-year medical students

Novices are generally advised to use a systematic search strategy for the evaluation of medical images. A systematic search strategy is defined as always evaluating images in the same specific order. Systematic search should lead to more complete evaluations (defined as coverage; the percentage of the image looked at), and fewer missed abnormalities. In a previous study, a systematic search strategy was not beneficial for the

detection of lesions by third-year medical students, when compared to a nonsystematic control condition. It is hypothesized that novices already need to possess some knowledge basis on cardiopulmonary diseases to take full advantage of a systematic search strategy for the various abnormalities on chest radiographs. Therefore, the effects of a systematic-viewing strategy compared to a nonsystematic viewing strategy on visual search patterns and lesion detection of chest radiographs in final-year medical students (n=60) were investigated with eye-tracking technology. Although students of the systematic-viewing group became more systematic compared to the students of the nonsystematic viewing group, both groups increased similarly in coverage and lesion detection post-training. Teachers are thus advised to particularly teach recognizing normal and abnormal findings on medical images and focus less on teaching systematic search strategies.

Chapter 4: The development of visual search patterns and lesion detection of first-year residents in radiology

Intermediates, such as residents in radiology, predominantly learn from feedback on their image evaluations as they engage in workplace learning. More insights into how the evaluation process takes place and how this process changes over time could provide additional and in-depth feedback to residents. However, previous investigations on the evaluation process in radiology generally were cross-sectional. Therefore, it is challenging to answer how the evaluation process of residents changes over time. The longitudinal development of visual search patterns and lesion detection on chest radiographs of first-year residents (n=16) in radiology was investigated with 11 experimental sessions consisting of 20 chest radiographs during the first year of residency training. Evaluation times were halved during the first four months. More efficient visual search patterns accompanied this decrease with the most pronounced changes happening during the first four months. Moreover, visual search patterns were slightly different on abnormal images compared to normal images. Finally, lesion detection remained constant throughout the first year of residency training, and any longitudinal effects of training on lesion detection were probably indirect as evaluation time did decrease. Overall, this study's findings provide more insights into how the evaluation process changes over time and could be used to enhance feedback for residents in radiology.

Chapter 5: The effects of evaluation order of traditional (plain) and new (contrast-enhanced) mammograms on visual search patterns and malignant lesion detection by breast radiologists

Acknowledged experts, such as radiologists, need to adapt and implement new imaging techniques into daily practice continuously. Contrast-enhanced mammography (CEM) is a recently introduced imaging technique and was found superior to full-field digital mammography (FFDM) for detecting malignant breast lesions. CEM examinations consist of a low-energy (LE, similar to a plain mammogram) image and a recombined, contrast-enhanced image (RC). Manufacturers typically advise LE-RC evaluation orders, yet breast radiologists with some experience in evaluating CEM report using an RC-LE order since malignant lesions appear more salient on the RC images. The effects of an RC-LE and a LE-RC evaluation order on visual search patterns and malignant lesion detection of breast radiologists (n=27) were investigated and compared to an FFDM condition. Evaluation times were 33% lower for the RC-LE order compared to the LE-RC order, while visual search patterns and lesion detection measures were similar. CEM conditions scored superior compared to FFDM on lesion detection, while evaluation times were similar. Eye-tracking technology proved beneficial to uncover a part of the previously covert evaluation process and is advised to tailor the implementation of new imaging techniques.

Chapter 6: General discussion

The main findings of the separate studies are summarized, and their theoretical and practical values are subsequently appraised. First, on the subject of teaching radiology, it is advised to focus more on the anatomy, physiology, and potential pathology, and less on systematic search strategies. Second, teachers should consider the proportion of normal and abnormal images of their image evaluation training. The prevalence of diseases impacts the criterion to differentiate normal from abnormal images. Teachers could use prevalence to shift novices' criterions. Third, eye-tracking technology can provide more insights into the image evaluation process for new and additional feedback to intermediates (residents) and recommendations to experts (experienced radiologists) for implementing new techniques into clinical practice. Furthermore, some limitations of this thesis need to be mentioned: only two-dimensional images were used in the experiments. Any extrapolation of the findings to volumetric images should be done with care. Additionally, the experiments on image evaluation training did not have delayed posttests to measure the retention of knowledge.

In conclusion, this Ph.D. thesis shows how different educational strategies are essential to support learners from the whole range of expertise development in radiology.

Samenvatting

S

Hoofdstuk 1: Algemene introductie

Het belang van medische beelden neemt toe in de dagelijkse, medische beslisvorming. Daarnaast beoordeelt een toenemend aantal (niet-radiologische) artsen deze beelden tegenwoordig zelf en er worden regelmatig nieuwe beeldvormende technieken geïntroduceerd. Het beoordelen van medische beelden wordt als een complexe vaardigheid beschouwd, en om hier deskundig in te worden is het noodzakelijk om uitgebreid te oefenen en ondergaat de leerling vele trainingen. Beginners, gevorderden en experts hebben andere leerervaringen gedurende de ontwikkeling van hun beoordelingsvaardigheden van medische beelden. Verschillende strategieën zijn dan ook essentieel om de leerervaringen van beginners, gevorderden en experts te ondersteunen. Het doel van dit proefschrift is om een leven lang leren te ondersteunen, door te onderzoeken hoe het leren beoordelen van medische beelden plaatsvindt van het hele spectrum aan artsen; van medisch studenten tot aan radiologen.

Beginners, zoals medisch studenten, hebben meestal hun eerste ervaringen met het beoordelen van medische beelden tijdens trainingen. Om beginners optimaal te ondersteunen, is het relevant om te onderzoeken hoe zulke trainingen zo effectief en efficiënt mogelijk kunnen zijn. De effecten van prevalentie van normale en abnormale beelden en het onderwijskundige (chronologische) ontwerp op de detectie en analyse van laesies door beginners, die röntgenfoto`s van de borstkas ("thoraxfoto`s") beoordelen, worden onderzocht in Hoofdstuk 2. De effecten van een training in systematisch kijken op visuele zoekpatronen en de detectie van laesies door beginners, die thoraxfoto`s beoordelen, worden onderzocht in Hoofdstuk 3.

Gevorderden, zoals arts-assistenten in de radiologie, houden zich bezig met werkplekleren. Zij leren hoofdzakelijk door feedback op hun eigen beoordelingen van de medische beelden. Meer inzichten hoe het beoordelingsproces plaatsheeft, en hoe dit proces ontwikkeld wordt, kan arts-assistenten voorzien van nieuwe en additionele feedback. Deze feedback kan gebruikt worden om meer inzage te krijgen in hun ontwikkeling. De ontwikkeling van visuele zoekpatronen en de detectie van laesies van eerstejaars arts-assistenten in opleiding tot radioloog, die thoraxfoto`s beoordelen, wordt onderzocht in Hoofdstuk 4.

Ten slotte, hoewel radiologen gezien worden als de experts in het beoordelen van medische beelden, zullen ook zij hun beoordelingsvaardigheden moeten

blijven ontwikkelen omdat er geregeld nieuwe, beeldvormende technieken geïntroduceerd worden. Radiologen vragen zich wellicht af hoe zij zulke nieuwe technieken zo optimaal mogelijk kunnen implementeren in hun dagelijkse, medische praktijk. De effecten van de beoordelingsvolgorde van traditionele (conventionele), nieuwe (contrast-versterkte) mammogrammen op visuele zoekpatronen en de detectie van maligne laesies door mammoradiologen worden onderzocht in Hoofdstuk 5.

Hoofdstuk 2: De effecten van prevalentie en onderwijskundig ontwerp op de detectie van laesies en analyse door derdejaars medisch studenten
Ten eerste, training in het beoordelen van medische beelden richt zich hoofdzakelijk op abnormale bevindingen, terwijl beelden in de dagelijkse, medische praktijk voornamelijk normaal zijn. Deze discrepantie tussen training en medische praktijk zou kunnen leiden tot een verkeerde indruk van studenten van de prevalentie van ziekten. Ten tweede, tijdens trainingen in het beoordelen van medische beelden krijgen beginners meestal eerst een uitleg van een expert voordat zij gaan oefenen (deductieve volgorde). Beginners zullen zo enkel leren om een voorgezegde oplossing toe te passen. Voor beginners wordt daarom oefening voorafgaand aan uitleg van een expert geadviseerd (inductieve volgorde). Een inductieve volgorde zou moeten leiden tot "productief falen"; studenten zullen zelf oplossingen moeten verzinnen en zullen hierin in eerste instantie falen, maar zij worden wel volledig ondergedompeld in het oplossen van dit probleem. Dit zou uiteindelijk moeten leiden tot een beter begrip van het probleem. In een 2x2 experiment tussen groepen van derdejaars medische studenten (n=103) zijn de effecten van prevalentie (70% normale beelden versus 30% normale beelden in een oefenfase) en onderwijskundige volgorde (inductieve versus deductieve volgorden), op de detectie en analyse van laesies op thoraxfoto`s onderzocht. Er werd een wisselwerking tussen de sensitiviteit en specificiteit van de medische studenten gevonden. Studenten die hoofdzakelijk met abnormale beelden oefenden, detecteerden meer laesies op de eindtoets (hogere sensitiviteit). Studenten die hoofdzakelijk met normale beelden oefenden, waren meer geneigd om normale beelden als normaal te beoordelen op de eindtoets (hogere specificiteit). Er werd daarnaast gevonden dat de studenten van de inductieve groepen een vergelijkbare detectie en analyse van laesies op de eindtoets hadden als de studenten van de deductieve groepen. De studenten van de inductieve groepen gebruikten onverwachts minder tijd per thoraxfoto gedurende de oefenfase, vergeleken met de studenten van de deductieve groepen.

Wellicht hebben de studenten van de inductieve groepen de beelden niet in voldoende diepte bestudeerd. Productief falen heeft waarschijnlijk ook plaatsgevonden in de deductieve groepen en is daardoor niet beperkt tot inductieve volgorden. Samenvattend, de proportie van abnormale en normale beelden zou een belangrijke overweging moeten zijn voor (het ontwerpen van) trainingen in het beoordelen van medische beelden, daarnaast worden deductieve volgorden geadviseerd voor radiologisch onderwijs aan medische studenten.

Hoofdstuk 3: De effecten van een training in systematisch kijken op visuele zoekpatronen en de detectie van laesies door laatstejaars medische studenten

Beginners wordt meestal geadviseerd om een systematische zoekstrategie te hanteren voor het beoordelen van medische beelden. Een systematische zoekstrategie is gedefinieerd als het altijd in dezelfde volgorde beoordelen van medische beelden. Het systematisch zoeken zou moeten leiden tot meer complete beoordelingen (gedefinieerd als dekking; het percentage van een medisch beeld dat bekeken is) en minder gemiste abnormaliteiten. In een eerdere studie bleek een systematische zoekstrategie niet bevorderlijk voor de detectie van laesies door derdejaars medische studenten, vergeleken met een onsystematische controlegroep. Er wordt verondersteld dat beginners al enige kennis van cardiopulmonale ziekten moeten bezitten om een systematische zoekstrategie volledig te kunnen benutten voor het scala aan abnormale bevindingen op thoraxfoto`s. De effecten van een training van een systematische zoekstrategie op visuele zoekpatronen en de detectie van laesies op thoraxfoto`s door laatstejaars medische studenten (n=60) werden vergeleken met de effecten van een training van een onsystematische zoekstrategie. Hoewel de studenten getraind in een systematische zoekstrategie systematischer werden, in vergelijking met de studenten getraind in een onsystematische zoekstrategie, namen de dekking en de detectie van laesies na de respectievelijke trainingen in eenzelfde mate toe. Docenten wordt daarom geadviseerd om voornamelijk aandacht te besteden aan het leren herkennen van normale en abnormale bevindingen, en minder aandacht te besteden aan het aanleren van systematische zoekstrategieën.

Hoofdstuk 4: De ontwikkeling van visuele zoekpatronen en de detectie van laesies door eerstejaars arts-assistenten in opleiding tot radioloog

Gevorderden, zoals arts-assistenten in opleiding tot radioloog, leren

hoofdzakelijk door feedback op hun eigen beoordelingen van medische beelden aangezien zij bezig zijn met werkplekleren. Meer inzichten over hoe het beoordelingsproces plaatsheeft, en hoe dit beoordelingsproces ontwikkeld wordt over tijd kan aanvullende en diepgaande feedback voor arts-assistenten opleveren. Echter, eerdere studies naar beoordelingsprocessen van medische beelden hadden nagenoeg allemaal een transversale onderzoeksopzet (cross sectioneel). Het is zodoende moeizaam om op basis van deze studies de vraag te beantwoorden hoe de longitudinale ontwikkeling van het beoordelingsproces van arts-assistenten plaatsheeft. De longitudinale ontwikkeling van visuele zoekpatronen en de detectie van laesies op thoraxfoto`s door eerstejaars arts-assistenten ($n$=16) in opleiding tot radioloog is onderzocht gedurende hun eerste opleidingsjaar, met behulp van 11 sessies bestaande uit telkens 20 thoraxfoto`s. De beoordelingstijd halveerde gedurende de eerste vier maanden. Efficiëntere visuele zoekpatronen vergezelden deze afname, waarbij de meest uitgesproken veranderingen eveneens gevonden werden tijdens de eerste vier maanden. Verder waren visuele zoekpatronen op abnormale medische beelden enigszins verschillend vergeleken met de zoekpatronen op normale beelden. Ten slotte, de detectie van laesies bleef constant gedurende het eerste opleidingsjaar, en mogelijke longitudinale effecten van de opleiding zijn waarschijnlijk indirect aangezien de beoordelingstijd afnam. Samenvattend, de resultaten van dit onderzoek bieden meer inzichten in de longitudinale ontwikkeling van het beoordelingsproces, en de resultaten kunnen gebruikt worden om feedback aan arts-assistenten te verbeteren.

Hoofdstuk 5: De effecten van beoordelingsvolgorde van traditionele (conventionele) en nieuwe (contrast-versterkte) mammogrammen op visuele zoekpatronen en de detectie van maligne laesies door mammoradiologen
Erkende experts, zoals radiologen, zullen zich continu moeten aanpassen en zullen nieuwe beeldvormende technieken moeten implementeren in hun dagelijks werk. Contrast-versterkte mammografie (CVM) is zo`n recent ontwikkelde, beeldvormende techniek. CVM blijkt superieur te zijn aan traditionele, conventionele mammografie (CM) voor de detectie van maligne laesies aan de borsten. CVM-onderzoeken bestaan uit een laag-energiek beeld (LE, vergelijkbaar aan een CM onderzoek) en een gerecombineerd, contrast-versterkt beeld (GC). Fabrikanten van CVM adviseren normaliter een LE-GC volgorde, terwijl mammoradiologen met enige ervaring met het beoordelen van CVM- onderzoeken aangeven om een GC-LE volgorde te gebruiken, aangezien maligne laesies meer in het

oog springen op de GC-beelden. De effecten van een GC-LE en een LE-GC volgorde op visuele zoekpatronen en de detectie van maligne laesies door mammoradiologen (n=27) is onderzocht en vergeleken met een CM-groep. De beoordelingstijd was 33% korter voor de GC-LE volgorde, vergeleken met de LE-GC volgorde, terwijl visuele zoekpatronen en de laesie detectie maten vergelijkbaar waren. CVM-groepen scoorden superieur vergeleken met de CM-groepen op de detectie van maligne laesies, terwijl de beoordelingstijden vergelijkbaar waren. Eye-tracking technologie bleek van meerwaarde om het beoordelingsproces beter in kaart te brengen. Deze technologie wordt geadviseerd voor de implementatie van nieuwe beeldvormende technieken.

Hoofdstuk 6: Algemene discussie

De algemene bevindingen van de afzonderlijke studies zijn samengevat, en de theoretische en praktische waarden zijn vervolgens afgewogen. Ten eerste, ten aanzien van trainingen om medische beelden te leren beoordelen, wordt geadviseerd om meer aandacht te besteden aan de anatomie, fysiologie en mogelijke pathologie, en minder aandacht aan systematische zoekstrategieën. Ten tweede, docenten zouden de proportie van normale en abnormale beelden in overweging moeten nemen voor hun trainingen in het beoordelen van medische beelden. De prevalentie van ziekten is van invloed op het criterium om normale bevindingen van abnormale te differentiëren. Docenten kunnen de prevalentie van ziekten daarmee gebruiken om dit criterium van beginners te verschuiven naar de meest wenselijke situatie. Ten derde, eye-tracking technologie kan meer inzichten bieden in het beoordelingsproces voor nieuwe en diepgaande feedback aan gevorderden (arts-assistenten). Daarnaast kan eye-tracking technologie aanbevelingen geven aan experts (ervaren radiologen) voor het implementeren van nieuwe technieken in de dagelijkse praktijk. Vervolgens worden enkele beperkingen van dit proefschrift benoemd: er zijn enkel tweedimensionale beelden gebruikt in de verschillende studies. Het veralgemeniseren van de onderzoeksresultaten naar volumetrische beelden (bijvoorbeeld MRI- en CT-scans) wordt hierdoor bemoeilijkt. Verder hadden de experimenten naar trainingen van het leren beoordelen van medische beelden geen uitgestelde eindtoetsen. Concluderend, dit proefschrift laat zien hoe verschillende, onderwijskundige strategieën noodzakelijk zijn om het hele spectrum van artsen (van medische studenten tot aan ervaren radiologen) te ondersteunen in het leren beoordelen van medische beelden.

Valorisation

This valorization addendum reflects on how the findings of this Ph.D. thesis on lifelong learning in Radiology can be utilized outside of the scientific field. First, the public relevance of this thesis will be discussed. Second, the main findings and potential strategies to support lifelong learning in radiology will be discussed. Third, the intended audience for the valorization of this thesis is described.

Public relevance

Many physicians are nowadays involved in the evaluation of medical images. Let us start with an everyday example of medical care in the Netherlands: A patient wakes up in the middle of the night, experiences serious dyspnea, and has developed a fever. He or she needs to go to the hospital. A clerk and a resident in internal medicine accommodate our patient. They suspect pneumonia and therefore request a chest radiograph to obtain more information about the current state of the patient's lungs and heart. Radiographs are generally considered complex to evaluate. The clerk and the resident evaluate the radiograph themselves, yet they struggle to come to conclusions. Therefore, they call the attending resident in radiology. The attending resident also evaluates the image and draws preliminary conclusions. Based on the preliminary conclusions, the resident in internal medicine starts a treatment plan. The next morning, a senior radiologist checks and finalizes the preliminary report of the resident. In this example, four (future) physicians are involved with only one radiograph. So how are all of these physicians trained to evaluate medical images?

The findings of this Ph.D. thesis indicate that all the physicians from the example have different training needs. Therefore, different strategies are necessary to support their learning experiences. The clerk, who may be considered a novice in evaluating medical images, probably received little training in evaluating chest radiographs. To support the learning experience of novices, it is necessary to focus on different aspects of current image evaluation training. The resident in our example can be considered an intermediate learner in the evaluation of medical images; he/she will have seen various normal and abnormal medical images already as they evaluate radiographs on a daily basis. Intermediates will particularly benefit from feedback on their own evaluations of medical images. For intermediates, it is therefore advised to provide in-depth feedback on their evaluations, as frequently as possible.

The radiologist of our example, who is an expert in evaluating chest radiographs, will have continued learning experiences throughout his/her career. New imaging techniques frequently become available, and experts will need to keep on adapting to an ever-changing medical field. For experts, additional support in implementing new imaging techniques into their everyday clinical practice is therefore advised.

Even though novices', intermediates', and experts' learning experiences substantially differ, all these physicians share the same, universal goal. They all strive to improve their image evaluations to provide the best possible care. By improving their learning experiences, their evaluations should subsequently improve. Improved evaluation skills should finally lead to fewer diagnostic errors and, thus, improved patient care.

Supporting learning of novices in radiology

For novices, many initiatives have already been developed for teaching medical image evaluation. Many medical students are trained with a lecture, followed by some practice cases and feedback to evaluate chest radiographs in a semester about cardiopulmonary diseases. Such lectures consist of basic cardiopulmonary anatomy, the radiological manifestations of some diseases, and instruction to always evaluate radiographs in the same, similar order, called a systematic viewing strategy. Our investigations show that there is a mismatch between the prevalence of diseases in image evaluation training and (future) medical practice: In image evaluation training, the prevalence of diseases is generally high, while most medical images in clinical practice are normal. This mismatch may impact the decision-making process of (future) physicians and may lead to more false-positive evaluations. Thus, a higher proportion of normal images in image evaluation training is advised.

Moreover, it is debated when to provide a practice phase with radiograph cases in image evaluation training. Medical students could practice before they receive an expert's explanation in a lecture, whereby they have to figure out solutions for themselves. Moreover, medical students could also practice after an expert's explanation, in order to check whether they understood this expert's explanation. Our findings show that for image evaluation training for medical students, the timing of practicing with radiograph cases does not matter for the detection of abnormal areas (lesions). Since a sequence with expert' explanation followed by practicing is

generally more time-efficient, this sequence is advised for medical students' image evaluation training.

Additionally, this thesis' findings indicate that teaching a systematic viewing strategy to medical students does not lead to increased detection of lesions on chest radiographs compared to a non-systematic (random) viewing strategy. Radiology teachers are therefore advised to focus on particularly the anatomy, the normal findings on radiographs, and the radiological manifestations of diseases instead of teaching viewing strategies.

Supporting learning of intermediates in radiology
Intermediates, such as radiology residents, could particularly benefit from additional and in-depth feedback on their own evaluations. Eye-tracking techniques capture where, when, and for how long a person has looked and could provide new and rich feedback to intermediates. In one investigation of this thesis, the first-year residents' eye movements were captured 11 times during their first year of residency training, while they were evaluating chest radiographs. One innovative aspect of a study on the longitudinal development of residents of this thesis is that the findings of this study may be used as a reference category of eye movement development: The eye movements of residents, while evaluating chest radiographs, could be regularly captured with eye-tracking technology during their first year of residency training. Their eye movement patterns can subsequently be compared to the eye-movement patterns of our longitudinal investigation. Such a comparison could provide residents with an additional source of feedback to monitor their development.

Additionally, eye-movement patterns could also provide in-depth information to residents on how their image evaluation takes place. Since eye-tracking technology is capable of capturing where, when, and for how long a person has looked, it can tell what specific areas of a radiograph residents have not laid their eyes upon. Such information on the evaluation process could be combined with information about whether residents missed lesions on that particular radiograph. Information about the evaluation process, with eye-tracking technology, and the outcome of the evaluation, such as missed lesions, could provide in-depth and more complete feedback to residents. Residents could use such information to improve their image evaluations.

Supporting lifelong learning of experts in radiology

Finally, eye-tracking methodology in medical image evaluation research has been primarily used to improve our understanding of how learning to evaluate images takes place. Another innovative aspect of this thesis is that one of our studies focused on senior radiologists learning to work with new imaging techniques. Recently a new imaging technique, contrast-enhanced mammography (CEM), has been introduced, which consists of a conventional radiograph of the breasts (mammogram) and a contrast-enhanced mammogram. CEM is superior for the detection of breast cancer lesions compared to conventional mammograms only. Radiologists were previously advised to evaluate the contrast-enhanced image after the conventional mammogram. Our investigation showed that an evaluation order with the contrast-enhanced image before the conventional mammogram led to similar detection rates of breast lesions, yet the evaluation was 33% more efficient compared to an evaluation order with the conventional mammogram followed by the contrast-enhanced mammogram. With eye-tracking methodology, it was unraveled that particularly the analysis of potential lesions was more efficient, not the detection of lesions. Overall, the use of eye-tracking methodology led to new insights into how this new imaging technique could be used most effectively and efficiently in everyday medical practice. Therefore, to implement new imaging techniques into experts' everyday medical practice, studies with eye-tracking methodology are advised.

V

Intended audience

In the last decades, radiological images have become widely accessible throughout hospital facilities through their digitalization. Therefore, many (future) physicians, from medical students to senior radiologists, evaluate radiological images on a daily basis. This thesis investigated this whole spectrum of learners in radiology. Therefore, the intended audience for the knowledge valorization of this thesis is physicians involved in the evaluation of medical images.

However, radiological images are not the only images that contain abundant information about the function and dysfunction of the human body. In our example, only a chest radiograph was obtained. In everyday medical practice, however, other medical tests are frequently ordered for an even more complete picture of the patient. One could think of electrocardiograms and laboratory blood tests, as well as pathology slides to be examined under a microscope by a pathologist.

Such medical tests, which also contain visualizations or representations of the human body, are also considered complex to evaluate. Therefore, the findings of this thesis on lifelong learning experiences in radiology could also apply to learning to evaluate these medical tests. It should be noted that pathology slices are generally evaluated by pathologists and residents in pathology only. For the field of pathology education, our findings on learning experiences may thus primarily apply to the spectrum from intermediates to experts. By contrast, other medical tests, such as electrocardiograms and laboratory blood tests, are generally evaluated by a broader range of physicians with different expertise levels. Therefore, the findings of our investigations can have added value for the whole range from novices to experts in learning to evaluate electrocardiograms and laboratory tests.

Dankwoord

D

DANKWOORD

Ellen, Simon en Jeroen. Dank jullie wel dat jullie mijn promotieteam wilden zijn. Alle drie afzonderlijk hadden jullie een duidelijke meerwaarde voor mijn ontwikkeling en deze thesis, maar juist de combinatie leidde tot een nog mooiere synergie van onderzoek, onderwijs en radiologie.

Ellen, als mijn dagelijks begeleider heb ik ongelooflijk veel geleerd van al onze discussies over expertise-ontwikkeling en eye-tracking. Ik bewonder je immense energie en oog voor detail op het gebied van onderzoek. Zelfs als alle andere coauteurs allang akkoord waren, zag jij mogelijkheden om een artikel nog scherper en argumenten nog sterker te maken. Je hebt me tijdens mijn PhD geprikkeld wanneer het kon en ondersteund wanneer ik dat nodig had. Ik hoop dat je nog lang je inzichten op ons gebiedje van onderwijs in de radiologie zal blijven delen.

Simon, je bent een geweldige raadgever en wegwijzer binnen de radiologie. Je eindeloze geduld en je netwerk hebben deuren geopend binnen alle centra waar ik deelnemers wilde werven. Ondanks je overvolle schema gingen we samen naar Eindhoven of Nijmegen om daar met de opleiders te praten over mogelijkheden tot participatie. Daarbovenop wist je mij er telkens weer aan te herinneren dat we dit onderzoek bovenal doen voor geïnteresseerde radiologen. Dit is de leesbaarheid van dit proefschrift voor de algemene radiologen bijzonder ten goede gekomen. Je bent voor mij een voorbeeld als radioloog en docent.

Jeroen, ik ben je enorm dankbaar voor de combinatie van vrijheid en sturing die je me hebt gegeven als promotor. Je gaf me de vrijheid om zelf richting te geven aan de verschillende studies, maar wist me telkens weer bij te sturen als ik dreigde af te dwalen. Ik bewonder daarnaast je analytische vermogen. Ik kan me bijvoorbeeld nog goed herinneren hoe je voorstelde om de volgorde van contrast-mammogrammen om te draaien, een nieuwe techniek waar je tot dan toe nog niet eens van gehoord had.

Prof. dr. Cees van der Vleuten, prof. dr. Anique de Bruin, dr. Linda Jacobi, prof. dr. Halszka Jarodzka en dr. Dirk Rutgers, graag dank ik jullie voor de tijd en moeite die jullie hebben gestoken in de beoordeling van dit proefschrift.

Lieve Karin, zonder jou was dit proefschrift nooit voltooid. Zonder jou had ik de gordiaanse knoop die ons gezinsleven, het restant van mijn promotieonderzoek, mijn werk als arts-assistent en onze hobby's waren geworden nooit kunnen ontwarren. Er is nu een duurzame stof van geweven, met jou en Julie als rode draad. Ik hou ontzettend veel van je en ik bewonder je discipline en doorzettingsvermogen. Je hebt ons een prachtige dochter geschonken en ik koester de dagen die we samen zijn. Dankjewel voor alles.

Lieve Julie, hoeveel ik ook lees over leren en onderwijs, van jou leer ik elke dag het meeste. Ik geniet van je nieuwsgierigheid, je eindeloze enthousiasme en bovenal je gulle glimlach. Hoeveel je ook elke dag verandert en jij je ontwikkelt, ik hoop dat je elke dag blijft stralen.

Lieve Mientje, Jan, Ingrid en Maarten, bedankt voor al jullie hulp tijdens de eindsprint, tijdens het schrijven van de thesis. Wat ik ook van jullie vroeg, het was nooit te veel gevraagd. Doordat Julie bij jullie regelmatig uit logeren kon, kon ik meters maken en kwam de finish razendsnel in zicht. Julie wordt eindeloos gelukkig van jullie en kan elke week niet wachten totdat het weer "donderdag oppasdag" is.

Tijdens mijn verdediging ben ik geflankeerd door een docent en een assistent in opleiding tot radioloog als paranimfen, een betere symboliek kan haast niet. Babs, je verhalen als docent Nederlands zijn altijd legendarisch en herkenbaar. Bovendien, je kan nog zoveel meer dan doceren; ik bewonder hoe je buiten de gebaande paden denkt en van je hobby, reizen, je beroep probeert te maken. Ivo, je bijdrage aan de studie naar contrast-mammografie was van onschatbare waarde. Zo redde je de dataverzameling door halsoverkop naar de ECR in Wenen te crossen, toen Karin onverwachts van Julie was bevallen. Het is fijn om samen met jou als vriend aan de opleiding tot radioloog in het Zuyderland begonnen te zijn!

Jorian and Abdullah, it was my pleasure to be your SCIP-internship supervisor. I learnt a lot through our supervision moments. Our discussions were helpful to sharpen my own thoughts about education, medicine and radiology. Jorian, my special thanks for driving with Ivo to the ECR of Vienna! Both of you were excellent students, and are now great colleagues.

Ik heb geweldig genoten van mijn PhD-tijd op de verschillende AIO-kamers.

We gingen naar Pubquizzen in de John Mullins, we spraken over Tina Turner, konijnen in scanners, hardloophorloges, Elton John, onze supervisoren, en soms zelfs over radiologische casus, of onderzoek naar onderwijs. Jorrick, Ellen, Katerina, Lorette, Andrea, Carolin, de Sannes, Stephanie, Rianne, Britt, Miriam, Serge, Samantha en Felicitas, allemaal onwijs bedankt voor alle leuke, gezellige, grappige en ontroerende momenten samen.

De vakgroep Radiologie van het Zuyderland Medisch centrum, en Wendy en Roy in het bijzonder, jullie gaven mij op het juiste moment het vertrouwen dat ik nodig had. Ik ben er trots op dat ik bij jullie mijn opleiding tot radioloog mag volgen en ik probeer elke dag het gekregen vertrouwen terug te betalen. Ik wil jullie daarnaast bedanken voor de tijd en ruimte die ik heb gehad voor het afronden van mijn proefschrift. Het creëren van deze ruimte is zeker niet makkelijk geweest in de tijd van onderbezetting en zomervakantie. Mijn dank is dan ook groot aan mijn collega-assistenten, Bram, Rik, Babs en Ivo, die het mogelijk hebben gemaakt dat ik tijdens de eindsprint vrijwel geen diensten had.

Simon, Marc en Ulrich, jullie hebben mij als radiologen en gouden standaarden geholpen met het maken van het juiste casusmateriaal voor de verschillende studies. Diederick, jij hebt alle eye-tracking bestanden van de longitudinale studie (circa 170 stuks) uitgewerkt tot de verschillende eye-tracking maten en diagnostische maten. Jimmie, Arno en Jeroen, bij jullie kon ik altijd terecht voor mijn statistische vraagstukken. Mijn dank is groot aan jullie allen!

Een gecombineerd promotietraject van radiologie en onderwijskunde, en deelnemers uit veel verschillende centra, betekent dat ik ondersteuning van vele secretariaten heb gekregen. Ik wil dan ook Lilian, Nicky, Audrey, Ryan, Monique, Christianne, Elfie, Joyce, Phyllis, Anja, Germien en Afke bedanken voor de talloze afspraken met deelnemers en begeleiders die zij in hebben weten te plannen!

Als laatste wil ik al mijn deelnemers bedanken voor de tijd die zij beschikbaar hebben gesteld voor dit onderzoek. Dit varieerde per deelnemer van 1 sessie van circa 30 minuten tot aan 11 sessies van minimaal 1 uur en jullie bijdrage is voor mijn onderzoek van grote waarde geweest. Zonder jullie bereidheid om deel te nemen en zonder jullie geduld om bijvoorbeeld voor de zoveelste keer de eye-tracker te kalibreren, was deze thesis er eveneens nooit geweest.

Curriculum Vitae

C

Koos van Geel was born in Breda on the 29th of March, 1989. He started medical school at the Maastricht University in 2007. He quickly became intrigued by medical education, and worked as a part-time research assistant of Maastricht University on various projects on educational research during his undergraduate education. Moreover, he took a gap year to advise the board of the Maastricht University Medical Center (MUMC+) on student matters as a student-assessor.

In his final year of medical school he completed a scientific internship on radiology education under the supervision of Prof. Dr. Simon Robben and Dr. Ellen Kok. He finished medical school in 2014, and after receiving the Kootstra Talent Fellowship of the MUMC+ he worked as a full-time Ph.D. student on radiology education at the School of Health Professions Education of Maastricht University. Since 2018, he is employed as a resident in radiology at the Zuyderland Medical Center in Heerlen.

List of Publications

Duvivier RJ, **van Geel K**, van Dalen J, Scherpbier AJJA, van der Vleuten CPM. Learning physical examination skills outside timetabled training sessions: what happens and why? Advances in health sciences education: theory and practice. 2012;17(3):339-55.

Cnossen TT, Konings CJ, Fagel WJ, van der Sande FM, **van Geel K**, Leunissen KM, et al. Fluid state and blood pressure control: no differences between APD and CAPD. ASAIO journal. 2012;58(2):132-6.

Gegenfurtner A, Kok EM, **van Geel K**, de Bruin A, Jarodzka H, Szulewski A, et al. The challenges of studying visual expertise in medical image diagnosis. Medical Education. 2017;51(1):97-104.

van Geel K, Kok EM, Dijkstra J, Robben SG, van Merrienboer JJ. Teaching Systematic Viewing to Final-Year Medical Students Improves Systematicity but Not Coverage or Detection of Radiologic Abnormalities. Journal of the American College of Radiology: JACR. 2017;14(2):235-41.

Kok EM, **van Geel K**, van Merriënboer JJG, Robben SGF. What We Do and Do Not Know about Teaching Medical Image Interpretation. Frontiers in Psychology. 2017;8(309).

Gegenfurtner A, Kok EM, **Van Geel K**, de Bruin AB, Sorger B. Neural correlates of visual perceptual expertise: Evidence from cognitive neuroscience using functional neuroimaging. Frontline Learning Research. 2017;5(3):14-30.

Kok EM, De Bruin AB, **van Geel K**, Gegenfurtner A, Heyligers I, Sorger B. The Neural Implementation of Surgical Expertise Within the Mirror-Neuron System: An fMRI Study. Frontiers in human neuroscience. 2018;12:291.

van Geel K, Kok EM, Aldekhayel AD, Robben SGF, van Merriënboer JJG. Chest X-ray evaluation training: impact of normal and abnormal image ratio and instructional sequence. Medical Education. 2019;53(2):153-64.

van Geel K, Kok EM, Krol JP, Houben IPL, Thibault FE, Pijnappel RM, et al. Reversal of the hanging protocol of Contrast Enhanced Mammography leads to similar diagnostic performance yet decreased reading times. Eur J Radiol. 2019;117:62-8.

SHE dissertation series

D

The SHE Dissertation Series publishes dissertations of PhD candidates from the School of Health Professions Education (SHE) who defended their PhD theses at Maastricht University. The most recent ones are listed below. For more information go to: https://she.mumc.maastrichtuniversity.nl

- Bourgeois-Law, G. (03-09-2020) Conceptualizations of remediation for practicing physicians
- Giuliani, M. (19-05-2020) A Critical Review of Global Curriculum Development, Content and Implementation in Oncology
- Schreurs, S. (20-03-2020) Selection for medical school; the quest for validity
- Schumacher, D. (19-03-2020) Resident Sensitive Quality Measures: Defining the Future of Patient-Focused Assessment
- Sehlbach, C. (21-02-2020) To be continued…. Supporting physicians' lifelong learning
- Kikukawa, M. (17-12-2019) The situated nature of validity: Exploring the cultural dependency of evaluating clinical teachers in Japan
- Kelly, M. (10-12-2019) Body of knowledge. An interpretive inquiry into touch in medical education
- Klein, D. (06-11-2019) The performance of medical record review as an instrument for measuring and improving patient safety
- Bollen, J. (01-11-2019) Organ donation after euthanasia: medical, legal and ethical considerations
- Wagner-Menghin, M. (25-09-2019) Self-regulated learning of history-taking: looking for predictive cues
- Wilby, K. (02-07-2019) When numbers become words: Assessors' processing of performance data within OSCEs
- Szulewski, A. (20-06-2019) Through the eyes of the physician: Expertise development in resuscitation medicine
- McGill, D. (29-05-2019) Supervisor competence as an assessor of medical trainees; Evaluating the validity and quality of supervisor assessments
- Van Rossum, T. (28-02-2019) Walking the tightrope of training and clinical service; The implementation of time variable medical training