Radiology Expertise in Diagnosing Frontal and Lateral Chest Radiographs: An Eye Tracking and Think Aloud Study

Fadi A. Toonsi

Department of Educational and Counselling Psychology

McGill University, Montreal

August 2018

A thesis submitted to McGill University

in partial fulfillment of the requirements of the degree of

Master of Arts in Educational Psychology

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Expertise is a well-known area of educational psychology discussed in a growing body of literature. Diagnostic radiology is one of many medical specialties interested in the development of expertise. Radiologists interpret images such as chest x-rays to come up with diagnoses by relating visual input to their background medical knowledge and associated patient information. Due to the visual nature of radiology, eye tracking has been an important tool to understand differences in expertise. Think aloud analysis is another methodology commonly used, as it provides a window into radiologists' thought processes. Pathways to expertise have been studied using these methods in radiology subspecialties such as mammography and chest x-rays. Although the latter is one of the most common radiologic investigations used in diagnostic radiology, a close examination of the expertise literature identifies two important gaps: (a) few studies examine expertise in reading lateral (side view) chest x-rays which, when well examined, provide important information; and (b) a dearth of research investigates types of false positive mistakes made by novices and experts, and any differences in the nature of these mistakes between these two groups of expertise.

This study addressed such a gap by examining how experts and novices differ in terms of their approach to reading frontal and lateral chest x-rays. Twelve expert and 11 novice radiologists were shown 21 normal and 21 abnormal frontal and lateral chest x-rays. Eye tracking and think aloud methods were used to examine differences between groups' diagnostic performance, decision time, false positive errors, and thought patterns. Findings reveal that experts were quicker and more accurate in their decisions compared to novices, while novices made more mistakes and took a longer time to decide. Different types of mistakes in radiology have different clinical significance, and our results demonstrate a clear difference between

groups in terms of the types of false positive mistakes made on both the frontal and the lateral chest x-rays. The most common false positive mistake made by novices was “blunting of the costophrenic angle,” whereas the most common false positive mistake made by experts was identifying a “lung opacity,” which may reflect experts’ emphasis on diagnosing serious conditions that frequently manifest as lung opacities. The lower performance and type of errors made by novices may reflect their lack of experience with both normal and abnormal cases, and the novice group’s comparatively limited knowledge of radiology.

Experts and novices enumerated a similar number of items during the think aloud process; however, when identifying a radiological finding, experts tended to actively search for related findings that could be present in similar clinical scenarios. For example, when compared to novices, more experts searched for evidence of axillary lymph node dissection after identifying a previous mastectomy. Such search for related findings may reflect greater in-depth knowledge of the mechanisms underlying the radiologic presentation of diseases and a deeper understanding of their various outcomes. This study highlights important differences between novices and experts in reading the chest x-ray and can aid in mapping out pathways to expertise in this area.

RÉSUMÉ

L'expertise est un domaine bien connu de la psychologie éducative discuté dans un corpus croissant de littérature. La radiologie diagnostique est l'une des nombreuses spécialités médicales intéressées par le développement de l'expertise. Les radiologues interprètent des images telles que les radiographies thoraciques (RT) pour arriver à des diagnostics en reliant la contribution visuelle à leurs connaissances médicales de base et les informations patient associées. En raison de la nature visuelle de la radiologie, le suivi oculaire a été un outil important pour comprendre les différences d'expertise. Penser à haute voix est une autre méthodologie couramment utilisée, car elle fournit une fenêtre sur les processus de pensée des radiologistes. Les voies d'accès à l'expertise ont été étudiées en utilisant ces méthodes dans des sous-spécialités de la radiologie comme la mammographie et les radiographies pulmonaires. Bien que ce dernier soit l'un des examens radiologiques les plus couramment utilisés en radiologie diagnostique, un examen attentif de la littérature spécialisée identifie deux lacunes importantes : a) peu d'études examinent l'expertise dans la lecture des radiographies latérales (vue de côté) qui peuvent fournir des informations importantes; et b) un manque de recherche étudie les types d'erreurs faussement positives commises par les novices et les experts, et toute différence dans la nature de ces erreurs entre ces deux groupes.

Cette étude a comblé une telle lacune en examinant comment les experts et les novices diffèrent dans leur approche de la lecture des RT frontales et latérales. Douze radiologues experts et 11 novices ont été montrés 21 RT frontales et latérales normales et anormales, respectivement. Des méthodes de suivi oculaire et de réflexion à haute voix ont été utilisées pour examiner les différences entre la performance diagnostique des groupes, le temps de décision, les fausses erreurs positives, et les schémas de pensée. Les résultats révèlent que les experts étaient plus

rapides et plus précis dans leurs décisions que les novices, tandis que les novices faisaient plus d'erreurs et prenaient plus de temps pour décider. Différents types d'erreurs en radiologie ont des significations cliniques et nos résultats démontrent une nette différence entre les groupes en termes de types d'erreurs faussement positives faites à la fois sur les RT frontales et latérales. L'erreur fausse positive la plus fréquente commise par les novices était « l'affaiblissement de l'angle costophrénique », tandis que l'erreur fausse positive la plus courante commise par les experts était « l'opacité pulmonaire », ce qui peut refléter l'accent mis par les experts sur le diagnostic de problèmes graves qui se manifestent fréquemment par des opacités pulmonaires. La baisse des performances et le type d'erreurs commises par les novices peuvent refléter leur manque d'expérience avec les cas normaux et anormaux, et leur connaissance relativement limitée de la radiologie.

Les experts et les novices ont énuméré un nombre similaire d'éléments pendant le processus de réflexion à voix haute; cependant, lors de l'identification d'une constatation radiologique, les experts avaient tendance à rechercher activement les constatations connexes qui pourraient être présentes dans des scénarios cliniques similaires. Par exemple, par rapport aux novices, un plus grand nombre d'experts ont recherché des signes de dissection ganglionnaire axillaire après avoir identifié une mastectomie antérieure. Cette recherche de résultats apparentés peut refléter une connaissance plus approfondie des mécanismes sous-jacents à la présentation radiologique des maladies et une meilleure compréhension de leurs différents résultats. Cette étude met en évidence des différences importantes entre les novices et les experts dans la lecture de la RT et peut aider à planifier un meilleur chemin vers l'expertise dans ce domaine.

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge and thank Dr. Susanne Lajoie, my supervisor. Her continuous support, enthusiasm, help, and guidance made this work more enjoyable and exciting.

This work might have not been done without the help of some wonderful family, friends, and colleagues and I feel obliged to sincerely thank each and every one of them. Thank you to my mother, my father, my wife, and my children for all the time, effort, and patience you provided while I was working on this thesis, and throughout the rest of my life-long learning journey. I expect the submission of this thesis ends the traditional part of that journey. What is next might be less demanding in terms of time and effort, but your care and support will always fuel my ambitions.

Furthermore, I would like to thank Dr. Louis-Martin Boucher for completing the scientific review of the research proposal. His comments and edits helped develop the study's design and improved its methodology.

In addition, I would like to thank Drs. Federico Discepola and Bojan Kovacina for their efforts in reviewing the experiment cases with me in order to establish a gold standard opinion on each one. Thank you for all those hours you volunteered to help me in this work.

I also would like to thank Dr. Dev Chakraborty from the University of Pennsylvania's Department of Radiology for volunteering his time and effort to guide me on ways to establish a proper methodology to test the performance of my participants. His efforts in Free Response Receiver Operator Characteristic are well known in the field of Radiology and it is my honour to have had his direct guidance on this research.

I also thank my sponsor, King Abdulaziz University, for funding my studies at McGill and providing me with the opportunity to pursue my goal of becoming an educator.

Finally, I thank my participants—all 23 of them—junior and senior residents, fellows and staff. If not for your time and effort this work would have never made it to the radiologists' light box.

PREFACE

I am the primary author of this Master's thesis and responsible for its content, from idea generation and literature review to ethics approval, data collection, statistical analysis, and manuscript writing. I wrote each of the chapters independently with my doctoral supervisor Dr. Susanne Lajoie reviewing and providing feedback at each step in the process of preparing the thesis. Dr. Lajoie supervised my work, supplied the required equipment to conduct the experiment, and provided scientific and logistical support required to write the manuscript. Dr. Lajoie also reviewed the manuscript prior to its submission. Dr. Dev Chakraborty contributed by helping design the experiment so it can maximally benefit from wAFROC statistical analysis. He also kindly conducted the wAFROC statistical analysis and provided its results. Dr. Wid Kattan and Mr. Mark Poulin proofread this thesis.

Results of this research project present original scholarship and are distinct contributions to knowledge. Specifically, the findings on radiology expertise demonstrate specific expert–novice differences in ability to diagnose and read chest radiographs. These findings were obtained through convergent methodologies, namely eye tracking and think alouds collected during the diagnosis process. Error detection and false positives were identified for each expertise group.

CHAPTER 1: INTRODUCTION

Diagnostic radiology imaging studies, including x-rays of the chest and abdomen, ultrasound examinations of various body parts, and magnetic resonance imaging, amongst others, stand in the front line of the investigative measures clinicians use to understand their patients' problems and plan their treatment. Once a set of images are acquired from a patient (also called an imaging study), they are available to physicians of other specialties to view and interpret. Consequently, interpretation of radiologic images is an essential part of almost all clinical training programs, such as general surgery and internal medicine. Diagnostic radiology training programs, however, are more in-depth, where future radiologists are trained to extract the maximal amount of information out of the images. Radiologists are the experts that other clinicians consult with when reviewing patients' imaging files.

Diagnostic radiology training is a long journey that involves a lot of learning, supervision, and dedicated practice. Trainees in diagnostic radiology training programs (also called radiology residents) spend their days in the hospital looking at radiologic imaging studies and interpreting them. Diagnostic radiology training involves a problem-solving approach and requires residents to gather as much pertinent information as possible about any given case. Trainees must synthesize information obtained from a variety of sources, namely the clinical information that is provided by the physician requesting the imaging study, the imaging study itself, and previous imaging studies the patient had prior to the current images taken. Trainees finally form a radiologic report that they feel best describes the abnormalities in the imaging study they are working on. They then review this report with a certified staff radiologist, discuss their findings, and edit the report before finalizing it. Trainees' learning process is segmented across these steps, and the road to becoming an expert requires continuous effort on multiple

levels. Novices learn something new on a daily basis, but due to the enormous amount of knowledge and experience required, the process takes a full 4 years of training before they are even eligible to enter certifying examinations for independent practice (J. Collins, 2001).

This training process transforms novice radiology trainees to experts in their fields who are rich with knowledge and ready to independently contribute to patient care. This novice-to-expert transformation is an important research subject that has been well-studied, and many publications describe the use of various research methodologies and investigative tools in this area. Many mysteries of this transformational path have been explored, but as with research in any field, an answer to one research question often opens doors to additional questions exploring beyond the original one.

This study aims to contribute to the existing body of knowledge by investigating the differences between novices and experts in reading frontal and lateral chest x-rays using reliable and well-established research methodologies: eye tracking and think aloud. One goal is to identify common patterns of performance gaps between experts and novices so that better instructional methods can be developed. Once these gaps are identified they can be used to provide novice learners with insights regarding their own performance and how it differs from that of experts. This two-step approach may lead to narrowing the performance gap between experts and novices in an innovative and efficient manner. Such understanding of how novices become experts can help in the creation and development of new teaching and learning tools that in turn could help speed up the educational process, thus enriching any remedial measures needed along the way.

CHAPTER 2: LITERATURE REVIEW

Expertise as a Field of Research

Cognitive research literature describes experts as individuals who excel in a specific domain of study and differ from a novice in that they make fast and accurate decisions, demonstrate reproducibly superior performance, are able to identify meaningful patterns, and have a better memory and higher self-regulation in their chosen field (Chi, Glaser & Farr, 1988; Ericsson, Krampe, & Teschromer, 1993; Johnson, 1988).

Psychologists have taken the lead in expertise research and set reliable paths for others to follow when conducting studies in this area. Their work includes expertise research in strategic games such as chess, as well as research on practical skills such as typewriting, programing, and mental calculations. Ill-defined problems such as judicial decisions and professional writing and expertise in arts and in sports have also been studied (Chi & Glaser, 1988; Chi, Glaser, & Farr, 1988; Ericsson, Charness, Feltovich, & Hoffman, 2006).

The long history of research on expertise reveals important characteristics of experts that appear consistently across various domains. These key characteristics serve as the base upon which current and future research stands. Following is a brief summary and discussion of these characteristics.

Expertise tends to be domain specific and experts usually stand out in their own fields. For example, highly rated chess players are experts in chess but when tested in a different game with which they are not familiar, they usually perform like novices. This characteristic is also noted in diagnostic radiology, and is important to keep in mind when interpreting results (Chi & Glaser, 1988; Nodine & Mello-Thoms, 2010).

Experts are also quick in finding accurate solutions for a problem they encounter. Such

speed stems from practice, familiarity with possible outcomes, and from greater domain knowledge (Chi & Glaser, 1988; Nodine & Mello-Thoms, 2010). In radiology, greater speed can result in seeing and reporting a larger volume of imaging studies, leading to higher efficiency of radiology services in hospitals; such characteristics are appealing to hospital administrations. Diagnostic radiology training programs tend to focus on developing their trainees' accuracy, knowledge, and performance more than they tend to focus on speed itself, knowing that if expertise is well developed then speed will follow.

Another characteristic of experts is that their analysis of problems is deeper than novices' and is based on domain specific principles. For example, when physics experts were asked to classify a group of problems, they used scientific principles for categorization; novices on the other hand categorized based on the literal types of objects they were presented with (Chi, Feltovich, & Glaser, 1981). Expert radiologists have deeper knowledge than novices in at least two domains: the physics of image formation and the mechanisms of disease. This knowledge informs their interpretation and categorization of radiological images. Novices on the other hand rely more on rote memorization of appearance of radiological signs without understanding of underlying mechanism (Manning, 2010; Nodine & Mello-Thoms, 2010).

Educators have expressed great interest in the study of expertise, and this is not surprising; becoming an expert is the ultimate goal of many educational programs, such as those in the field of medicine. From an educator's point of view, studying expertise can help identify a path leading toward becoming an expert in a specific domain, and help set up specific achievable and measurable landmarks on that path. A novice following through these serial landmarks eventually would become an expert (Lajoie, 2003).

When it comes to conducting experiments on expertise, there are two main approaches:

They can either be conducted in a lab setting where confounding factors are better controlled, or they can take place in a natural setting, where outcomes better reflect real-life scenarios. Expertise in both settings has been studied using various methodologies. Direct observation is one method, where experimenters record their observations, with the intent of analyzing how experts approach various tasks to determine the key processes that are used to solve problems. Methods such as conversational analysis and video-based interaction analysis are used to understand the different levels of expertise present in an observed experiment. These analyses involve cataloging the video data and analyzing participants' gaze and gestures. Two methodologies have gained momentum in studies of radiology expertise, namely, eye tracking of participants' gaze patterns across images, and verbal protocols that are used during the reading and interpretation of images.

Anders K. Ericsson, a prominent psychologist in this field, advised studying expertise using a structured "expert performance approach" (Ericsson & Smith, 1991). This approach comprises three steps. First, the task to be studied is developed into a reproducible experiment that the investigator will test repeatedly—in the lab or field. Then, the experimenter examines the methods that yield expert performance using process-tracing measures, such as eye tracking, verbal report, or even functional magnetic resonance imaging. Finally, researchers try to understand how these methods are acquired, so that they can be set as training targets for novices (Ericsson & Smith, 1991).

Research on radiology expertise fits quite well within the framework of the expert performance approach, perhaps even more so than some other clinical specialties, because it is easy to establish a reproducible experiment with a set of diagnostic images. In radiology, the lab setting is relatively similar to real life. Radiologists in both settings will take a seat in front of a

computer monitor and read diagnostic imaging studies on them. Since these images are portable and easily representable, unchanged, time and time again, the lab setting can faithfully represent the natural setting in terms of the problem-solving task: the diagnostic interpretation of radiology images. The stationary condition of the image reporting sessions makes it easy to study performance measures using different methods such as stationary eye trackers, which have higher accuracy compared to portable ones. Similarly, an interview-like protocol analysis method is easier to apply in a stationary lab setting, compared to a mobile dynamic setting of a medical ward for example. It is not surprising therefore to find research on expertise in diagnostic radiology dating back to 1975 (Kundel & Nodine, 1975).

Our study is based on the expert performance approach and aims to shed light on expert performance in a small but important division of radiologic studies; namely, the chest x-ray.

Expertise in Diagnostic Radiology

Expertise in medicine has been well studied. Indeed, a search in the PubMed medical literature database reveals more than 3,500 articles on expertise in medicine. These publications span across many medical specialties and radiology is not an exception. The study of expertise in radiology has been subjected to a good amount of research using various methodologies.

To understand expertise in radiology, one must define the scope being examined. A key objective for radiologists is to interpret imaging studies. This process involves two main activities: (a) processing perceived visual information, and (b) making decisions about them. Various factors play a role in these two tasks, and below is a summary that focuses on factors deemed relevant to this research.

Processing visual information in medical images involves understanding what specific constituents mean in the image. All the lines, shadows, and colours in an image must be

interpreted. X-ray images are perceived as representations of physical objects (patients' anatomy) and proper interpretation requires accurate corroboration of what is seen in the x-ray and what was previously learned about human anatomy. This interpretation requires an understanding of the physical properties of x-rays and their interaction with the components of the human body (Manning, 2010). In addition to understanding the general meanings of an image, trainees must differentiate between normal and abnormal image features. This second step in image interpretation demands decision-making about what is normal and what is not. Most radiology learning resources address the above components; they explicitly explain imaging findings, correlate them with physiologic abnormalities, provide different case examples, and discuss the making of diagnostic decisions based on specific imaging features (Nodine et al., 1999). It can be said therefore, that radiology experts excel at both visual processing and making better decisions based on what they see, and that continuous deliberate practice and case exposure leads to better expertise.

Before concluding this overview of expertise in diagnostic radiology, it is important to provide a brief remark on the role of deliberate practice. In order to become experts, radiologists must *deliberately practice* their skills. Ericsson et al. (1993) describe deliberate practice as long hours of repeated effortful activities designed specifically to improve a given skill. They argue that deliberate practice is more important than any innate abilities and that continuous focused training accompanied by careful and purposeful supervision leads to higher performance. The long hours of thought-provoking discussions with staff radiologists during reading review sessions provide the training residents with an excellent opportunity to deliberately practice and refine their image interpretation skills, directly supervised by experts.

After discussing research in expertise and expertise in radiology, the following section

addresses the more specific area of methodologies used to study expertise in radiology. As mentioned earlier, eye tracking and verbal reports—including think aloud protocols—are process-tracing measures that are commonly used to study expertise in radiology.

Eye Tracking Studies on Radiology Expertise

Eye tracking devices track participants' gaze while they view targets set by researchers. These devices rely on infrared cameras that focus on subjects' eye pupils and triangulate their central vision to identify what they are focusing on. In the early days of radiology, the target used to be the radiographic film, which changed in the late 1990s and early 2000s to diagnostic computer monitors. Computer software can be connected to the eye tracking camera that provides gaze measures. These measures include how long participants look at the image in general (total dwell time), how long they focus on a specific point in the image (target dwell time), the time it takes them to fixate their eyes on a target (time to target), and the pathways their eyes follow while viewing the image (scan/search pattern), amongst other measures (Nodine & Krupinski, 1998). By comparing experts' eye tracking metrics to those of novices, researchers can gain deeper insights between the two, thus gaining insight about the path towards competency and expertise.

Eye tracking has been extensively utilized to understand radiology expertise and previous studies have helped answer several questions related to how radiologists view and interpret imaging studies. Following is a summary of some key articles, selected from the large body of literature, that serve as landmark studies in the field of radiology expertise and highlight important gaps between novices and experts.

One of the first eye tracking studies on radiologists was conducted by Llewellyn-Thomas and Lansdown (1963). They studied radiology trainees when reading chest x-rays and found that

trainees' search patterns were not universal but rather trainee dependent. As for the area covered in the x-rays, interestingly, they found that different trainees tended to look at various parts of the image, while ignoring others. Another early study by Kundel and La Follette (1972) assessed the visual search patterns of normal and abnormal radiographs, comparing radiologists to radiology residents, medical students, and laymen. In contrast to Llewellyn-Thomas and Lansdown's findings on trainees, Kundel and La Follette found that radiologists showed specific patterns when viewing these images; they started with the hilar regions (more central area) followed by an assessment of the lateral aspects of the chest (peripheral areas). These patterns were interestingly quite different from the ones found in untrained subjects.

Nodine and Krupinski (1998) sought to determine, using eye tracking, whether radiologists' detection skills were domain specific; that is, whether they were good at detecting findings in radiological images in particular, or in images in general. Their experiment explored if the lesion detection abilities of radiologists extend outside their field and if they were any better than laypeople in detecting comic target figures (i.e., Waldo) on pictorial scenes. Radiologists in their study were no better than the public in detection performance. Radiologists actually took longer than average to first fixate their gazes on the target and spent a longer time to search the images. This implies that radiologists are not generally better at visual detection but rather excel in their field due to specific cognitive and visual skills that are related to their medical training, pointing to the importance of task-specific training. These findings are also seen in another study (Kelly, Rainford, McEntee, & Kavanagh, 2018) providing concurrent evidence to the domain-specificity features of expertise identified in other specialties (Chi & Glaser, 1988; Ericsson, 2006a ; Nodine & Mello-Thoms, 2010).

Eye tracking has been used to help design and assess educational interventions in

radiology, as well as in other medical specialties. In pediatrics, for example, Jarodzka et al. (2010) recorded gaze patterns of experts while they were examining infants with seizures. They then developed videos of these gaze patterns to teach medical students about this disease. These videos highlighted the parts of the patient's body that the expert was looking at, thus guiding learners to also focus on these areas. They reported enhanced diagnostic performance of epileptic seizures by the experimental medical student group (those who viewed the videos with attention guidance).

In a study that aimed to assess the effectiveness of a training intervention, Manning, Ethell, Donovan, and Crawford (2006) used eye tracking to assess the visual search strategies for lung nodules in radiographers (health allied professionals who perform imaging studies in patients but do not interpret them) both pre and post training, in comparison to radiologists and radiography students. Their training intervention improved radiographers' overall detection performance and changed their visual search strategies so that they replicate those of expert radiologists. They also found that radiologists used larger saccades (eye movements) across the image and tended to cover less of the area of the films. Interestingly, this was not associated with higher error rates.

The list of publications and educational interventions that have utilized eye tracking is long, but again one area where research appears to be small is the nature of misdiagnosis and errors novices make in comparison to experts. Eye tracking experiments can help identify common mistakes made by novices and experts alike. Our research used eye tracking as one method to localize error as a path to understand expertise. Another well-utilized method of studying expertise is the think aloud method, a type of verbal report analysis. This will be discussed below.

Think Aloud Studies on Radiology Expertise

Among the different types of verbal reports, the think aloud method is commonly used and involves recording experts' spoken thoughts while they perform the task in question. The recordings are then transcribed and coded using specific coding schemes. The coded transcripts are subsequently analyzed quantitatively and qualitatively in an effort to specify key elements of experts' performance (Ericsson, 2006b; Gegenfurtner & Seppanen, 2013). Similar to eye tracking methods, think alouds provide a measure of the thinking process as they seek to understand the mechanisms driving the problem-solving task. The think aloud method differs from other types of verbal report analysis in that there is minimal interference in participants' processes by the researchers during the data collection session.

Think aloud methods have been used in the medical field in general and in the field of radiology specifically. They have been used, for example, to identify perceptual and reasoning components of novice, intermediate, and expert pathologists while they view and diagnose breast histopathology slides—a visual task similar to what is seen in radiology (Crowley, Naus, Stewart, & Friedman, 2003). Results of this research were used to develop a cognitive model of competency in the pathology specialty.

Similar work was done by Azevedo, Faremo, and Lajoie (2007) but in radiology. They analyzed verbal reports of mammogram interpretations from radiologists, radiology residents, surgical residents, and medical students to develop a mammography problem-solving model and to characterize the differences between novices and experts in this diagnostic task. Their results were used to develop a teaching tool: a computer-based learning environment targeted at radiology trainees.

Gegenfurtner and Seppanen (2013) combined eye tracking and think aloud methodology to examine the transfer of expertise from familiar to semi-familiar and unfamiliar imaging

technologies. Their research shed some light on the concept of expertise transfer in radiology, as it showed that expert performance was transferred from the familiar imaging modality to the semi-familiar one, but not to the unfamiliar one.

Again, similar to eye tracking, not much research has investigated experts and novices approach to the lateral chest x-ray using the think aloud method and our work aims to help fill this knowledge gap.

The Lateral View of the Chest

Plain radiographs of the chest are amongst the most abundant radiographic studies done in hospitals. Their low radiation dose, ease of acquisition, and relatively high yield makes them an initial investigation physicians frequently request to assess their patients' torso region.

Routine plain chest radiographs are acquired mainly in two projections: the postero-anterior projection (PA, a view from the back of the patient) and the lateral projection (a profile view from the side of the patient). Other projections are available on demand, but are somewhat less frequently requested.

The PA projection is considered the primary projection in chest radiography, while the lateral is less frequently requested. Each projection, however, as with all plain radiographs, has an inherent limitation in its assessment of depth; while the x-ray beam passes through the patient's body, it summates the shadows of all the tissue it passes through in its direction and forms a composite image on the film. This summation limits depth assessment on radiographs and hinders three-dimensional localization of lesions. In other words, all the 3-D data is "squeezed" together to form a 2-D image. To overcome this limitation and to localize findings on the missing third dimension, a perpendicular projection can be acquired. In the case of chest imaging, the lateral projection is the perpendicular projection that complements assessing the

depth of a normal structure identified on the PA view.

The lateral view provides rich and necessary information in many aspects. It clarifies normal structures that overlap with one another on the PA projection, such as the mediastinum and the great vessels. It also clears up anatomical areas that are obscured on the PA view, such as the posterior sulcus of the lung and the retrosternal area. Although some of this information can be indirectly inferred from the PA view using subtle radiological signs, the lateral film quickly and directly provides this information. Certain structures, such as the spine, are simply best assessed on the lateral view as it stands clear from overlying structures (Gaber, McGavin, & Wells, 2005).

When it comes to abnormalities, the lateral view can help physicians assess the depth of a lesion that is visualized on the PA view, as well as detect hidden lesions within the chest. Masses and nodules in areas such as the mediastinum and around the lung hila can be confirmed on the lateral view. The lateral view also clarifies lesions obscured by the bony structures such as the sternum and the vertebral column. A pneumo-mediastinum (air within the mediastinal pleural cavity) is better assessed on the lateral view than on the PA. The lateral can also help identify subtle rib fractures in trauma cases (Robinson, 1998).

In addition, the lateral view can help separate normal anatomical structures from pathologies, as in the case of a normal lung vessel seen end-on on the frontal view, deceptively appearing as a pathological lung nodule or a granuloma. Finally, the lateral projection can help confirm or reject any suspicion of an abnormality seen on the frontal view, such as in the cases of right middle lobe collapse, or fissure abnormalities.

Sagel, Evens, Forrest, and Bramson (1974) expressed the importance of acquiring a lateral radiograph with almost all PA views. They argued that the only time it could be omitted was when a chest radiograph is obtained for screening purposes in otherwise healthy people who

are under 40 years of age. Otherwise, Sagel et al. note, each PA should be accompanied by a lateral view. Many other authors emphasized this point time and time again (Delrue et al., 2011; Feigin, 2010; Robinson, 1998)

Current Clinical and Teaching Trends on the Lateral View

Despite the useful information provided by the lateral view, it is underutilized and seldom requested by clinicians—that is, physicians who are in direct contact with patients in clinics and emergency rooms. New-generation clinicians do not always request a lateral view when a frontal view is obtained; if they do request it, they tend to under-study it. To them, the PA view is usually the first and, often, the last view looked at. Any information obtained from the PA view is (mistakenly) considered the maximum "practical" amount of information obtainable from the plain film modality, and clinicians' next step is often to request more complex and costly radiological exams such as CT scans or MRIs (Feigin, 2010; Gaber et al., 2005; Robinson, 1998). In contrast to clinicians, radiologists continue to value the lateral view and understand the role it can play in patient care. However, it is the clinicians and not radiologists who request radiological studies. Clinicians also usually interpret plain films on their own once their patients get them while waiting for radiologists' formal reports. Because clinicians are ordering fewer lateral views, radiologists are seeing and reporting fewer as well.

There are different explanations for the trend of overlooking the lateral film by clinicians, and most revolve around the expertise required for its interpretation. The normal anatomy is more difficult to interpret on the lateral film because in it, a large three-dimensional volume of the body is projected into a small two-dimensional film, leading to an overlap of the radiographic shadows of multiple organs on top of each other. The lateral film thus requires deeper knowledge of the normal anatomy of the chest, and of abnormal radiological signs. Therefore, a greater

amount of time and effort is spent on interpreting a lateral radiograph. Longer interpretation times decrease the functionality of the lateral projection; thus, a busy clinician tends to jump to CT scans which provide more information, despite the risk of exposing the patient to higher doses of potentially harmful radiation. The low demand for the lateral film leads to less exposure to this kind of radiological exam, which in turn leads to less experience and expertise on the lateral view. Lower expertise again drives clinicians away from requesting, and hence a *negative expertise loop* ensues. Low demand for the lateral film also affects radiologists because it means they will do less reporting of and will get less experience with the lateral view. Therefore, extra care should be given to compensate and enhance radiologists' training.

Considering the decreasing role of the lateral view in the era of cross-sectional imaging as discussed above, it is not surprising that diagnostic radiology, as a specialty, emphasizes the significance of the lateral film. Interpreting lateral images is given a high importance by the older generation of radiologists, and is an art that is passed with care from teaching staff to residents in-training. In an article titled "The Lateral Chest Radiograph: Is It Doomed to Extinction?" Robinson (1998) highlighted the importance of reviewing the training methods used to teach the lateral view and pointed out that its interpretation requires additional educational efforts.

Newer techniques and teaching methods are published every now and then, and lectures and presentations dedicated to teaching lateral film interpretation are often seen in radiology conferences. Feigin (2010), for example, published an article describing a systematic approach to reading a lateral chest x-ray, one that takes into consideration the CT scan era we are currently in.

Only a few published articles address experts' approach to multiple projections of one region of the body. For example, Calvin Nodine et al. (1999) studied the identification of mammographic abnormalities on two views, and found that higher expertise yielded better

reporting and classification of paired lesions, i.e. lesions appearing on both projections. To our knowledge, there are no eye tracking studies investigating how trainees and experts look at both the frontal and lateral plain films of the chest when presented together, nor are there studies examining individuals' thought processes using the think aloud method while they interpret these types of imaging studies. There appears to be a knowledge gap in the literature on the expertise on the lateral film, and that is the target of our study.

Errors in Diagnostic Radiology

Diagnostic errors are an important cause of patient distress and account for up to 75% of malpractice claims against radiologists (Lee, Nagy, Weaver, & Newman-Toker, 2013). One important reason for studying radiology expertise is to identify the types of errors made by each group, which in turn can help build a trajectory toward becoming an expert in this field (Lajoie, 2003). There are four main sources of errors in diagnostic radiology identified in the literature: observer errors, recognition errors, interpretation errors, and communication errors (Pinto et al., 2012). Observer errors refer to when the radiologist fails to look at the abnormality. Recognition errors refer to when the radiologist looks at the abnormality but fails to recognize it as abnormal. Interpretation errors involve providing wrong interpretations and explanations of radiologic findings. False positives (the interpretation of a normal finding as abnormal) may fall into this category. Finally, there are communication errors, in which radiologists fail to provide their diagnostic opinion in an appropriate and timely manner (Pinto & Brunese, 2010). An overall error rate of radiologists has not been agreed upon, but the number is believed to be around 30% for the average radiologist. Chest radiographs show a similar number as well. These figures might appear high, but are well supported by research, especially that which explores malpractice lawsuits (Berlin, 1986, 1996). Most literature on error investigates false negative

errors and the number above is concerned with that type of error. Not much literature looked into false positive errors, and their rates remain to be investigated.

Research on malpractice provides a wealth of information regarding the most common types of radiological mistakes (Berlin & Berlin, 1995; Whang, Baker, Patel, Luk, & Castro, 2013). However, such research focuses mostly on practicing radiologists rather than trainees. Research on errors peculiar to trainees is scanty and not well established.

It is important to note that key literature on errors in radiology mostly focuses on false negatives rather than false positives. A false negative, which can also be called a "missed diagnosis," is an abnormal finding on a radiological image that has not been detected by the radiologist. A false positive occurs, on the other hand, when a radiologist reports an abnormal finding that in fact does not exist—for example, when a radiologist reports the presence of a fracture in a normal bone (Kok et al., 2016; Manning, Ethell, & Donovan, 2004; Whang et al., 2013). The greater focus on false negatives is probably due to the significant and morbid consequences of delayed diagnosis of diseases such as cancer. Studies on false positives are less frequent, and studies on trainees' false positives are even more scarce despite their importance. Raising a false positive concern about a potential disease can have negative effects on patients and their families, and lead to stress and unnecessary expensive and potentially harmful investigations. One of the merits of our study is that it addresses this particular gap in the literature, by allowing us to explore the common false positives trainees and experts report as they fixate their gaze and think aloud.

Relevance

The introduction in chapter 1 shows that expertise in reading chest x-rays is an important topic. Researchers have examined radiological expertise from different perspectives, however there still is a necessity for additional research. Literature about expertise in diagnostic

interpretations of the lateral view of the chest is sparse. Furthermore, little research addresses false positive mistakes made by experts and novices. The generation of experts that used to understand all the ins and outs of the lateral view is getting older, and the younger generation requires different training techniques in an era where CT scans prevail. A study that explores how experts perform diagnostic interpretations of lateral views and compares them to novices could narrow the gap between the two groups. Knowing how experts think and function can be used to guide novices in setting their own learning goals. Understanding the types of errors novices make can be useful in designing training strategies and teaching methods to help them attain higher levels of expertise. Knowledge of what novices and experts look at, how they progress over the years, and the way they develop deeper interpretations and richer thoughts about the lateral view can open important doors to better instruction in the future. For example, experts' "mental check lists" and "if–then" strategies can be used as objective goals that can inform teaching, learning and assessment. Collectively, these outcomes can guide radiology training programs on how to establish competency based training.

This study poses the following research question: Do expert radiologists perform better than novices on the postero-anterior (PA) and the lateral chest x-rays? Accuracy measures as well as process measures of performance (eye tracking and think alouds) are used to determine group differences. Specifically, we are looking for differences between the groups in their comfort level reading chest examinations, differences in their diagnostic performance indicators, in the time to decide on normality of cases, on the types of false positive mistakes made, and differences in the think aloud verbal analysis. The prediction is that experts will outperform novices in all these measures.

A mixed methods quasi-experimental study is used to assess performance differences and compare novice radiology trainees to expert radiology trainees and staff radiologists on PA and

lateral chest x-rays. The design is a mixed methods design as it utilizes both quantitative and qualitative data in the form of eye tracking metrics and think alouds, and it is quasi-experimental as participants are pre-assigned to one of two groups based on their expertise. The next chapter describes the methodology used to investigate these differences in more detail.

CHAPTER 3: METHODOLOGY

Participants

The target population was radiology residents, chest radiology fellows, and staff chest radiologists working at a diagnostic radiology department in a North American university (fellows are board certified radiologists taking additional training in a radiology subspecialty). Thirty-eight of these potential subjects were residents, two were chest fellows, and six were staff chest radiologists.

A $50 incentive was advertised and given to a randomly selected trainee from the resident pool. Participants signed an informed consent before data collection and were told they could withdraw from the study at any time they wished. Neither trainees' evaluators nor anyone other than the investigators knew how each individual participant performed.

Recruitment was voluntary and was mainly done by in-person discussion about the study, its goals, and its potential benefits. An invitation email was also sent to all radiology residents by the radiology department's secretary (a total of 38 residents received the email). Eighteen residents volunteered to participate (five were 1st-year radiology residents, six were 2nd-year residents, and seven were 3rd-year residents). Two fellows and five staff were invited verbally. Both fellows and three staff agreed to participate. Overall, the total number of participants was 23 (N = 23). The mean age of participants was 31.5 years (M = 31.5 years, SD = 5.3 years). Sixteen of the participants were males and seven were females. Radiology residency training in North America consists of 4 years of training. Radiology residents in their 3rd year onwards

were considered experts based on the common consideration of 3rd-year residents as “seniors” who have achieved basic competency and are awarded more autonomy and responsibility.

Of these participants, 11 were novices and 12 were experts. The novice group consisted of five 1st-year and six 2nd-year radiology residents. The expert group had seven 3rd-year radiology residents, two chest radiology fellows, and three staff. Both fellows finished a previous year of fellowship training in a subspecialty other than chest radiology. Figure 1 shows the training level of the participants of each group.

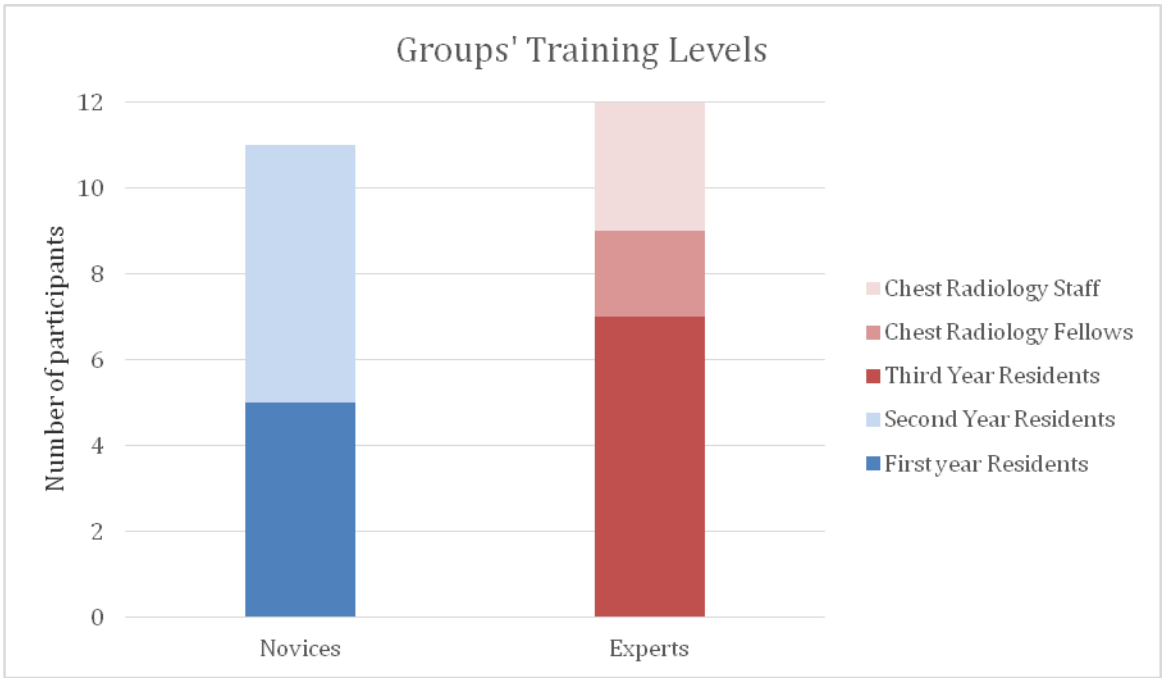


Figure 1. Level of training of the two groups of expertise.

There were four females and seven males in the novice group, while the experts group had three female and nine male participants. Figure 2 shows participants’ gender distribution for both groups. The groups were different in terms of age. The mean age of novices was 29.7 years (*M* = 29.7 years, *SD* = 4.8 years). The mean age of experts was 33.2 years (*M* = 33.2 years, *SD* =

5.5 years). Figure 3 shows the box and whisker plot of participants' age, separated by gender.

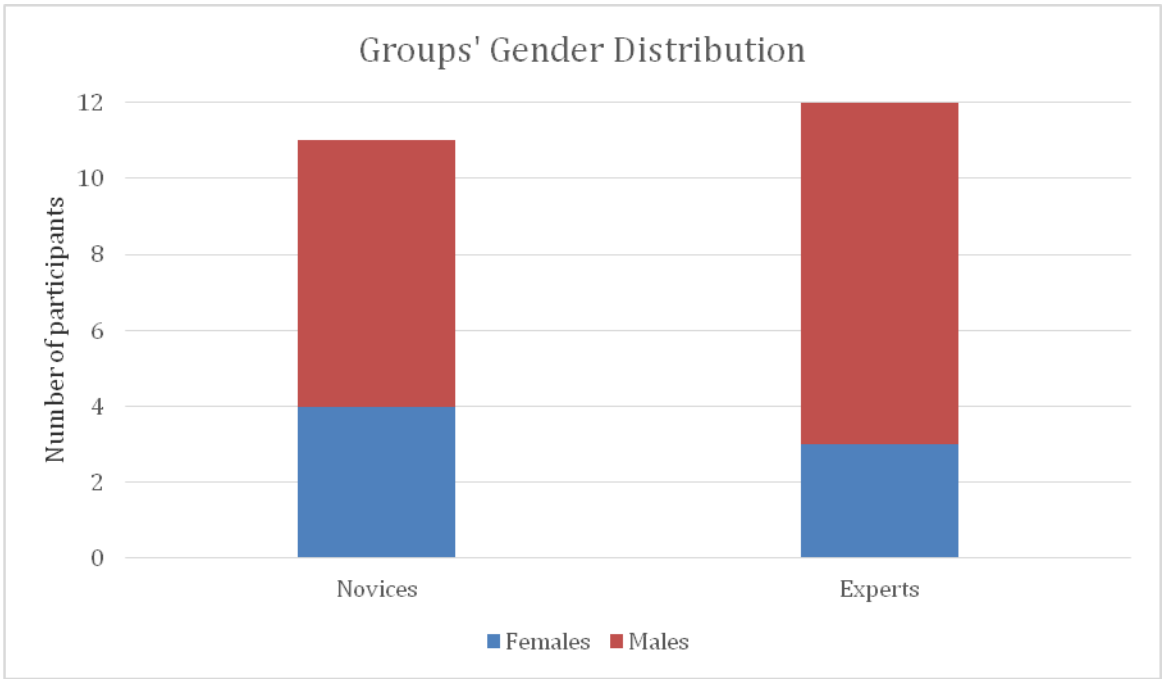


Figure 2. Groups' gender distribution.

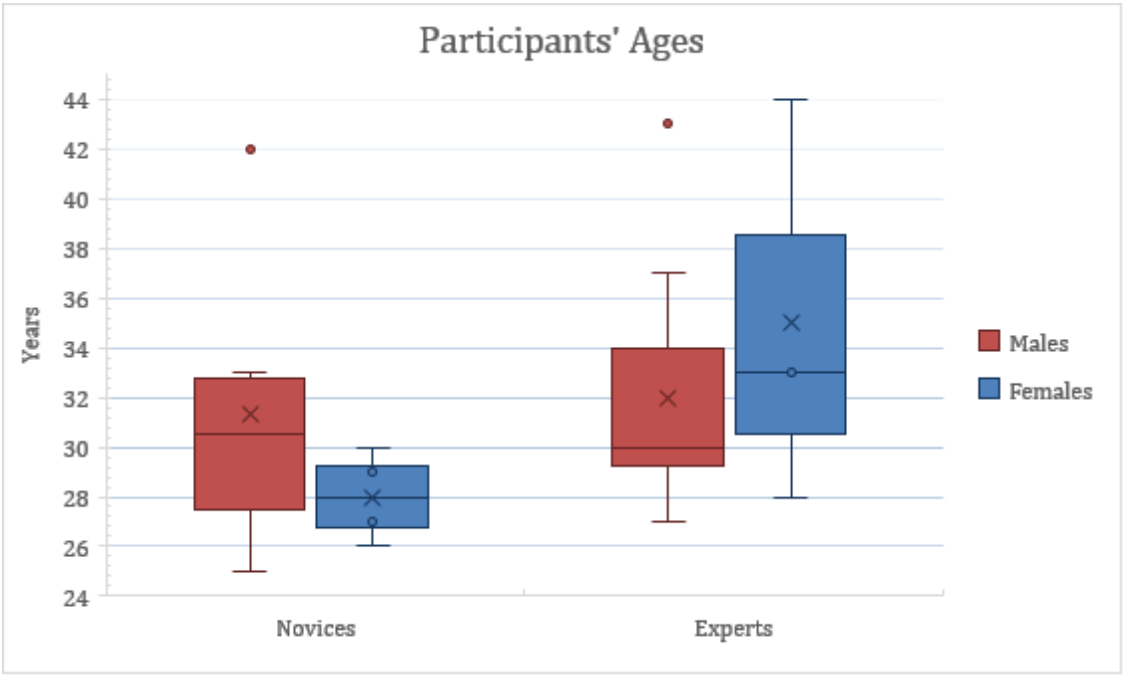


Figure 3. Participants' age box and whisker plot separated by gender.

Information about the use of eye glasses and contact lenses was collected since eye glasses, but not contact lenses, can potentially decrease accuracy of eye tracking data. Twelve participants (52%) did not require eye sight correction; four (17%) required eye glasses and were using them during the experiment; five participants (22%) required corrective measures and were using contact lenses during the experiment; and two (9%) required eye glasses but were not using them during the experiment. The latter two participants mentioned that they were comfortable to continue the experiment without their glasses and after reviewing their mistakes at the end of the experiment, mentioned that their mistakes were unlikely related to not wearing their glasses. Both participants were in the expert group.

Apparatus and Materials

Selection of Chest X-rays

A total of 42 cases were selected from the electronic medical records of a University Hospital. Each case had a frontal and a lateral view. Twenty-one of these were normal and 21 were abnormal. The 42 cases were split, so that 40 were used to collect eye tracking data and two were used to collect think aloud data.

The cases were chosen in the following manner: hundreds of cases were initially reviewed by the primary investigator[1], along with each case's previous and follow-up imaging studies and their reports, when available. The cases were selected so that participants would be exposed to a variety of lesions that differed in terms of anatomical location and pathology. The majority represented common findings, but a few cases with rare findings were also selected to maximize performance variability and help distinguish experts from novices. Some normal cases with potentially perceivable false positives were included as well, for the same reason.

[1] The primary investigator of this thesis is a radiologist with 2 years of experience.

After this initial screening process, the best 141 cases were selected for review by an expert panel consisting of three diagnostic radiologists: the primary investigator and two other practicing radiologists, one of whom also has additional training in chest imaging as a subspecialty. During panel review the experts were requested to point out all the abnormalities they found in both the frontal and the lateral views of each case, and to provide a difficulty level ranking (from 1 to 5) for each finding they identified. This difficulty ranking helped determine the difficulty level of the selected cases as well as ensure an overall balanced distribution across difficulty levels.

If the panel did not find any abnormalities in a case while reviewing, then that case was considered normal. Cases where members of the expert panel reported discrepant findings were either eliminated from the experiment pool or discussed again after gaining additional clinical information. Expert panel review sessions were conducted several times until a consensus between all the panel members was reached on 21 normal and 21 abnormal cases out of the original 141. These 42 cases were used in the experiment.

Equipment and Software

Selected films were imported into Adobe Photoshop CC 2015.5 (Version 17.02, Adobe Systems Software Ireland Ltd) where the PA and lateral films of each patient were placed next to each other. No resizing was done in this step. The PA film was always placed on the left side of the final image and the lateral was on the right. Unifying the location of each projection throughout the experiment rather than randomizing it would make participants more comfortable with the location of each projection and would minimize the time they would spend searching for a specific projection when presented with a new case. Measured times in this case would reflect actual thinking and diagnosing time rather than time spent searching for a projection. The PA

and lateral were combined to an image sized 4044 x 2022 pixels. When displayed on the experiment monitor this was downscaled to the monitor's native resolution, which is 1920 x 1200 pixels. Downscaling was done automatically using the eye tracking software.

All cases were presented on a Dell Ultrasharp U2413 23.8" monitor (Dell Inc.), borrowed from the department of diagnostic radiology. This monitor model is used in the department to view chest x-rays on a daily basis. This equipment was chosen because it provides acceptable quality for diagnostic purposes, while at the same time is portable enough to be moved to the different experiment rooms. The monitor was connected to the experimenter's laptop using a standard HDMI-DVI connection. Zooming, windowing, and panning were not permitted during the experiment in order to keep the time measured reflective of the actual detection task rather than to manipulation of these tools.

The eye tracking hardware consisted of a portable remote eye tracking infrared camera model RED manufactured by SMI (SensoMotoric Instruments GmbH) running at 120 HZ and placed below the participant's viewing monitor. This was connected to a laptop running the camera-control software iView X (Version 2.8 build 26, SensoMotoric Instruments GmbH). The laptop also ran SMI's Experiment Center software (Version 3.7 build 60, Professional 360° license, SensoMotoric Instruments GmbH) that was responsible for conducting the experiment by displaying the instructions to participants and showing the test chest x-rays in a randomized pattern. The software recorded the gaze metrics, viewing time, and mouse clicks. Experiment analysis was conducted on SMI's BeGaze software (Version 3.7 build 42, Professional 360° license, SensoMotoric Instruments GmbH).

The experiment was audio recorded from beginning to end for each participant using a portable voice recorder (Sony Corporation of America). A back-up recording was also obtained

using an apple iPhone 6S (Apple Inc.). The recordings were transferred to the investigator's laptop and deleted from the devices.

Measures

The following five measures were targeted to investigate the differences between participants' expertise:

1. A questionnaire that includes demographic information, experience, and comfort level on reporting chest x-rays (PA and lateral). (See Appendix: Electronic Questionnaire.)
2. Diagnostic performance, as measured by wAFROC statistical test, discussed below.
3. Eye tracking measures, specifically the total time to decide on normality of cases and visual search patterns.
4. False positive lesion localization analysis.
5. Think aloud verbal analysis.

The questionnaire was created for the purpose of this study and included questions about participants': demographics (fellowship or residency level, number of years of employment for staff physicians, age, gender, and the need and current use of eye glasses and contact lenses); area of expected and or previous radiologic subspecialty training; experience in radiology (in general); the number of chest rotations done to date; and the main reading source about chest radiology. Three questions asked about comfort level on reading frontal and lateral chest x-rays and chest CT-scans on a 1 to 5 Likert scale.

Diagnostic performance was measured by a statistic called wAFROC which takes into account both accuracy and confidence of participants as they examine each radiological image. wAFROC will be discussed below, but it is important to mention here that its calculation depends on the number of lesion localizations (true positives), non-lesion localizations (false positives), and confidence levels made by each participant. A lesion was considered correctly localized if the

participant accurately clicked on it on the x-ray film and correctly described its radiological features. The investigator would relate the location and description to what the truth panel agreed on and, based on that, mark the participant's mouse click as a lesion localization or a non-lesion localization. If a participant did not describe the lesion radiologically then the investigator prompted a description. Non-lesion localization (false positives) consisted of all the mouse clicks or radiological descriptions that did not correspond to a lesion agreed on by the truth panel.

Eye tracking measures were collected from 40 cases. Two more cases were used for think aloud data collection. The eye tracking measures that were collected were the total dwell time to decision for each case, and the visual search patterns.

False positive lesion localizations were identified by reviewing each participant's gaze, voice recording and mouse clicks. Each abnormality they called was localized to either the PA or the lateral view and was transcribed on an Excel sheet, classified as a lesion localization (true positives) or non-lesion localization (false positives). Non-lesion localizations were grouped, quantified, and analyzed.

Think alouds were transcribed. For the first case (the normal chest x-ray), anatomical structures mentioned by each participant were identified and quantified.

The second think aloud case was complex, with more than one finding on both the PA and lateral views. The transcriptions were examined for the presence and absence of specific findings. Additionally, the transcriptions were examined for the presence of any associated thoughts (i.e., patterns where certain thoughts lead to others). Different reasoning patterns between the groups were also explored (Fonteyn, Kuipers, & Grobe, 1993).

Experiment Setting

The experiment was conducted within the department of diagnostic radiology at two North American Hospitals. Due to variable availability, the experiment was conducted in one of

four rooms. The rooms were quiet and no interruptions occurred once an experiment started.

Each participant was assessed individually. They were seated comfortably in front of the eye tracking system. The experiment and its goals were explained, but no further emphasis was given to the two views to minimize potential bias. After they read and signed the consent, the audio recording and the eye tracking software were activated. Participants were instructed to use a computer mouse to click on radiological abnormalities and a keyboard to advance through the experiment slides. Participants' seat position was adjusted to a specific distance from the monitor, between 65 and 100 cm, as recommended by the manufacturer of the eye tracking device. An initial calibration screen was displayed at the beginning of the experiment and calibration was repeated when necessary.

Procedure

The study was reviewed by the Research Ethics Office of the Faculty of Medicine at the North American University and an approval was obtained beforehand. All the data collected from participants were anonymized. The chest x-rays had no patient identifiers when they were presented to reviewers and to participants. These test cases were reviewed for the purpose of this experiment only, with no intent to double read or change any previous radiological reports.

First, participants answered the questionnaire by filling an electronic form. After that the investigator explained the case review process and started the experiment. As mentioned, 40 cases were dedicated to collecting eye tracking measures. Half of these cases were normal and the other half were abnormal. After finishing the eye tracking part, the two think aloud cases were displayed. One of these two cases was normal and the other one was abnormal. All participants viewed the same 40 eye tracking cases and the same two think aloud cases. The order of cases was randomized between participants to minimize the practice effect bias. The

think aloud always followed the eye tracking part.

Each participant went through three demonstration cases to get used to the experimental procedure. These cases were not included in the analysis. The process of case reporting was as follows: It started with a screen that only displays the case number. When ready, the participant advanced the experiment to the next slide, which displayed a patient's frontal and lateral views next to each other. The total dwell time to decide whether the case was normal or not was calculated here for each case, as the participant was asked (beforehand) to press the space bar to advance to the decision slide as soon as they had made their decision. They were assured that they would have another chance to scrutinize the films for as long as they wanted to later on, in order to identify abnormalities.

Participants then advanced to the next slide which showed the same case for the second time. This time they were requested to click on and report all radiological abnormalities, and to click on a rating scale displayed on top of the x-ray to rate their confidence about each abnormality. The scale is from 1 to 5, with 5 being most confident. This confidence rating is commonly used for diagnostic performance analysis using wAFROC statistic. The participants were told that a radiological abnormality was any deviation from normal that lies inside the patient, this is to exclude ECG leads and film markers, etc. Participants were told to disregard any technical film related issues. Once they completed reporting all abnormalities, if any, participants clicked to advance to the screen that contained the next case's number.

Participants were told that each case could have single or multiple radiological abnormalities, or be completely normal. They were not aware of how many cases were normal and how many were abnormal.

In order to study participants' performance on the lateral film, they were asked to click on

their findings (and corresponding confidence levels) on the lateral view independently of the frontal view. In other words, they were asked to repeat a given finding detected previously on the frontal view if applicable, and rate their confidence about it on that view.

After going through the 40 cases, the participants then viewed the two think aloud cases. These two cases were randomized in their sequence of presentation to minimize potential bias. The participants were requested to verbalize their thought process while diagnosing the case. If a participant remained silent for more than a few seconds, they were asked to verbalize their thought process (Fonteyn et al., 1993). If at the end, the participant did not verbalize a final diagnosis of the case, they were asked to provide one; otherwise, there was not much interaction between the investigator and the participant during the think aloud section of the experiment. It should be noted that the think aloud cases were not included in the wAFROC analysis as previous studies showed that think aloud in general improves performance (Ericsson, 2006b).

All participants were shown the same cases and no time constraints were provided throughout the experiment. Once a participant finished the experiment, the voice recording was turned off and they were given the opportunity to discuss their performance with the experimenter as time permitted.

Statistical Analysis

The data were exported from the BeGaze software to Microsoft Excel (Microsoft Office 365 ProPlus, Version 1705, Build 8201.2200, Microsoft Corporation), where it was coded. It was then exported to SPSS (IBM SPSS Statistics, Version 24, Release 24.0.0.0. IBM Corp.). Descriptive statistics, Chi square and t-tests, were conducted in SPSS. The names of statistical tests are displayed accordingly in the results section. An α of .05 was used for significance testing throughout the experiment.

Measuring Competency in Radiology and the Receiver Operator Characteristic Curve

Participants' diagnostic performance was assessed using a measure called the Figure of Merit (FOM), popularized by Chakraborty and Berbaum (2004). The FOM is an outcome measure that uniquely combines a participant's accuracy and confidence to yield a score that reflects these components of expertise. It is important to note that "confidence" does not reflect a participant's confidence in their own abilities but rather their confidence on the accuracy of the finding they are reporting. Thus, confidence will vary from case to case even for the same participant. An example may be necessary to explain: experts may question an ambiguous shadow on a normal chest x-ray that they are doubtful means anything at all, yet choose to report it and admit their low confidence that it is a true finding. In this case, their low confidence will give them a higher FOM score in comparison to a novice who reports the same ambiguous shadow with a high confidence.

These scores can be used to compare radiologists' diagnostic performance. A higher FOM number indicates that participants made less false positive errors while also being less confident that they truly represent errors, and also made more true positives while being highly confident that they truly represent radiological abnormalities. The FOM is calculated by a statistical test called the Weighted Alternative Free-response Receiver Operator Characteristic (wAFROC) (Chakraborty, 2010, 2011; D. P. Chakraborty & Berbaum, 2004; Hillis, 2010; Metz, 1978; Tourassi, 2010). After calculating an FOM for each participant in each case (including both the frontal and the lateral views), a mean FOM is given to each participant. This reflects the participant's overall performance on all the cases, combining their performance on the frontal and the lateral views. The mean FOM of each group is then calculated, before comparing both groups.

wAFROC analysis, including FOM calculation, was conducted in R (a language and environment for statistical computing; R Foundation for Statistical Computing) using the wAFROC method described by Chakraborty and Zhai (2016). wAFROC analysis includes FOM calculation for each participant and testing the significance of the difference between the means of the two expertise groups by bootstrap analysis and z score testing. The wAFROC analysis was done as a random reader fixed cases analysis. wAFROC analysis was kindly provided by Dr. Dev Chakraborty. Further details and instructions on this method are available on his website.[2]

[2] http://www.devchakraborty.com/

CHAPTER 4: RESULTS

Our analyses were conducted to determine if and how expert radiologists perform better than novices on the postero-anterior (PA) and the lateral chest x-rays. In the results section we first describe the characteristics of the experts and novices in terms of demographic information followed by specific analyses on accuracy measures and process measures of performance (eye tracking and think alouds) .

Questionnaire Responses

The first area the questionnaire tapped into was participants' experience level. This was evaluated by inquiring about two variables: (a) years of experience in radiology, and (b) number of chest rotations. It was expected that experts would have more experience than novices. Experience in radiology was measured by the reported number of years of radiology training, added to the number of years working as a radiology staff. Experience in chest radiology was measured by the reported number of chest radiology rotations. A chest radiology rotation is defined as a 4-week period dedicated to reading and reporting cardiothoracic radiology, regardless if this period was during residency training, fellowship, or while practicing radiology afterwards.

Novices, expectedly, had less experience in radiology in general and in chest radiology in particular, compared to experts. They reported a mean of 2.6 years of experience in radiology (M = 2.6 years, SD = .5 years), and a mean of 1.9 chest rotations to date (M = 1.9 rotations, SD = .8 rotations). Experts on the other hand reported a mean of 6.8 years of experience in radiology (M = 6.8 years, SD = 4.2 years), and a mean of 26.8 chest rotations (M = 26.8 rotations, SD = 46.4 rotations). The difference between the groups was statistically significant for the number of years of experience in radiology, t (11.4) = -3.42, p = .005, but not for the number of chest rotations t (11) = -1.86, p = .091. Figure 4 displays the number of years of experience in radiology for the

participants of each group. Figure 5 displays the number of chest rotations for the participants of each group.

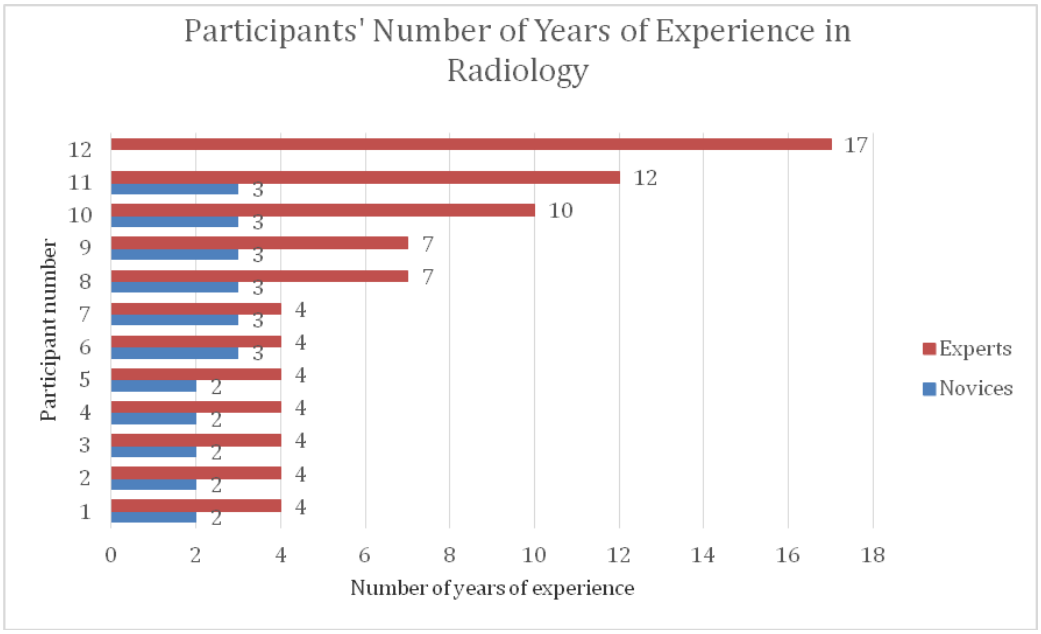


Figure 4. Participants' number of years of experience in radiology. For comparative purposes participants are arranged from lower experience at the bottom to higher at the top.

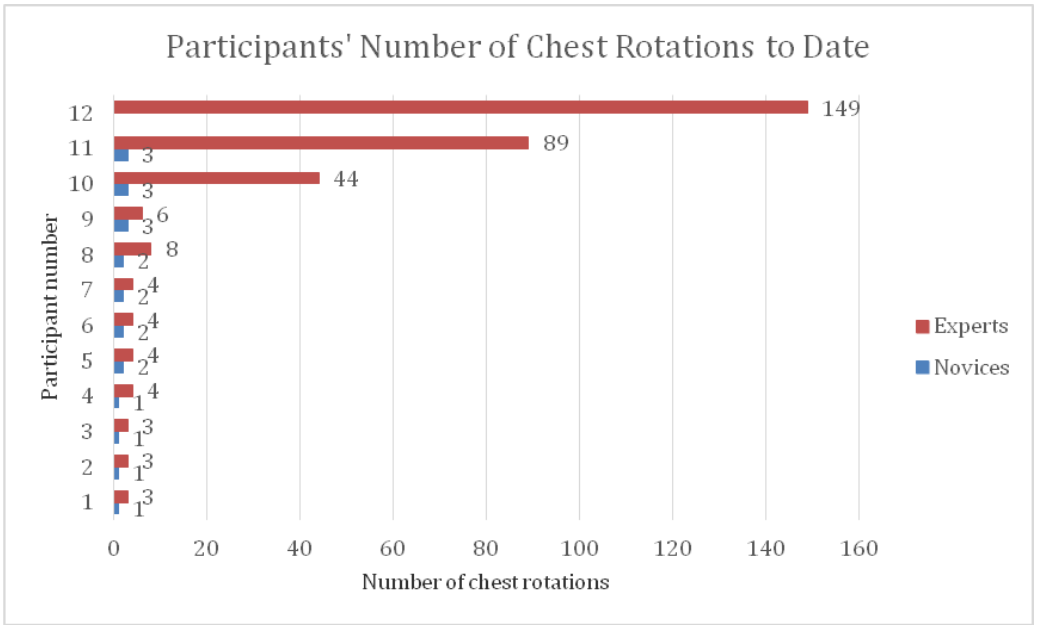


Figure 5. Participants' number of chest rotations to date. For comparative purposes participants are arranged from lower number of rotations at the bottom to higher at the top.

When asked about intended, current, or previous fields of subspecialty training, none of the novices reported intent to specialize in cardiothoracic radiology. Five experts were either training or previously trained in cardiothoracic radiology. Two of these were fellows and three were staff.

For participants' ranking of their comfort level reading chest x-rays, the Likert scale responses were given values from 1 to 5, with 1 corresponding to "very uncomfortable" and 5 to "very comfortable." When asked about comfort reading PA chest x-rays, the mean reported value for novices was 3.55. (M = 3.55, SD = .67). Experts on the other hand reported higher values (M = 4.17, SD = .58). The difference between the groups was statistically significant, t (21) = -2.35, p = .028.

When asked to rank their comfort level reading the lateral view, novices reported a mean comfort level of 3.3 (M = 3.3, SD = .65). Experts reported close values (M = 3.8, SD = .97). The difference between the groups here, however, was not statistically significant, t (21) = -1.38, p = .182.

As for their comfort level reading chest CT scans, novices reported a mean comfort level of 4.1 (M = 4.1, SD = .3). Experts reported a higher value (M = 4.75, SD = .45). The difference between the groups was statistically significant, t (19.28) = -4.14, p = .001. Figure 6 displays the participants' comfort levels when reading PA and lateral chest x-rays and CT scans of the chest.

The main source for reading about chest radiology varied. A spectrum of responses (Figure 7) indicated 12 different resources that can be grouped into two different categories: (a) general radiology resources, and (b) specialized cardiothoracic radiology resources. Ten novices and three experts reported a general radiology resource, while nine experts and a single novice reported reading mostly from a specialized cardiothoracic resource. There was a statistically significant relationship between expertise level and the category of the resource used for reading about chest radiology, where novices read mostly from general radiology resources while experts

read from specialized cardiothoracic radiology resources, $\chi 2$ (1, N = 23) = 10.14, p = .001.

Figure 8 shows the reported radiology resources grouped by category.

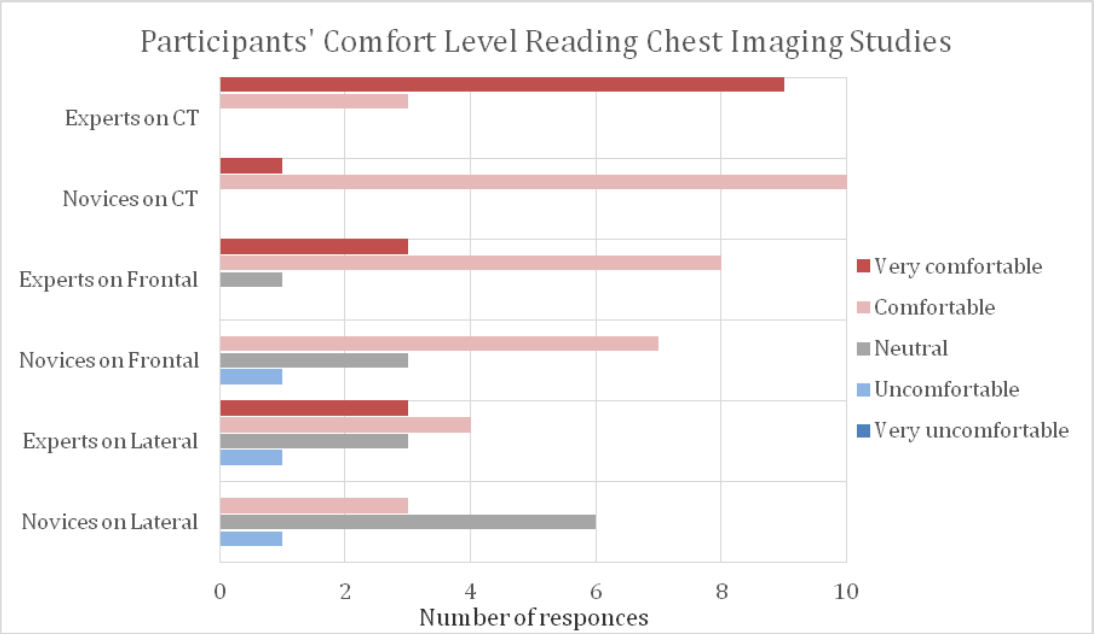


Figure 6. Participants' comfort level reading chest imaging studies.

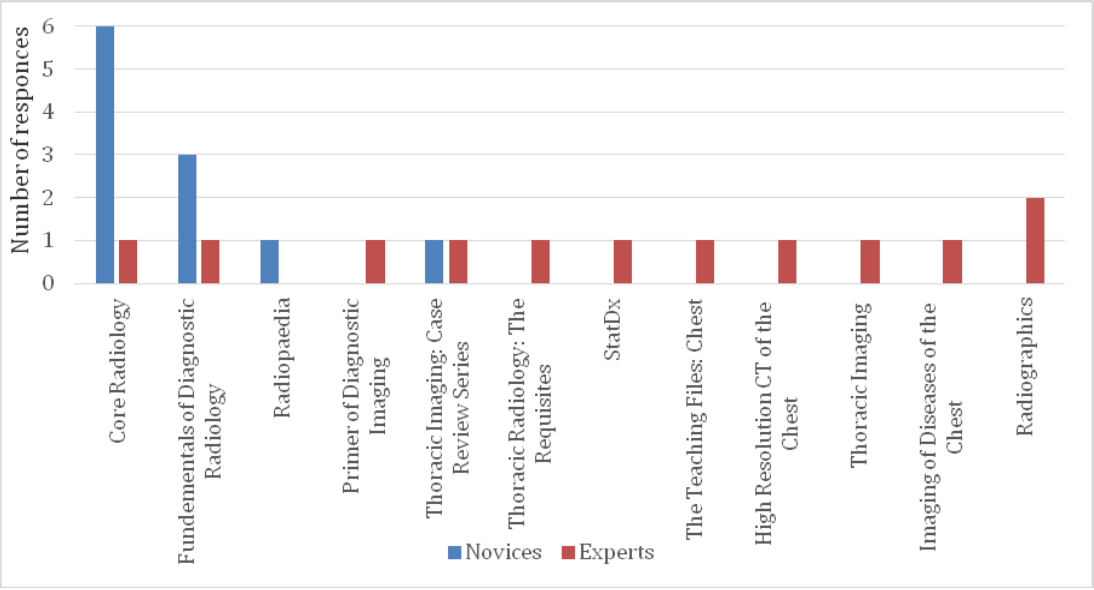


Figure 7. Participants' main source of reading for chest radiology. The first four references in the left of the figure can be grouped as general radiology sources. The remaining can be grouped as specialized cardiothoracic radiology sources.

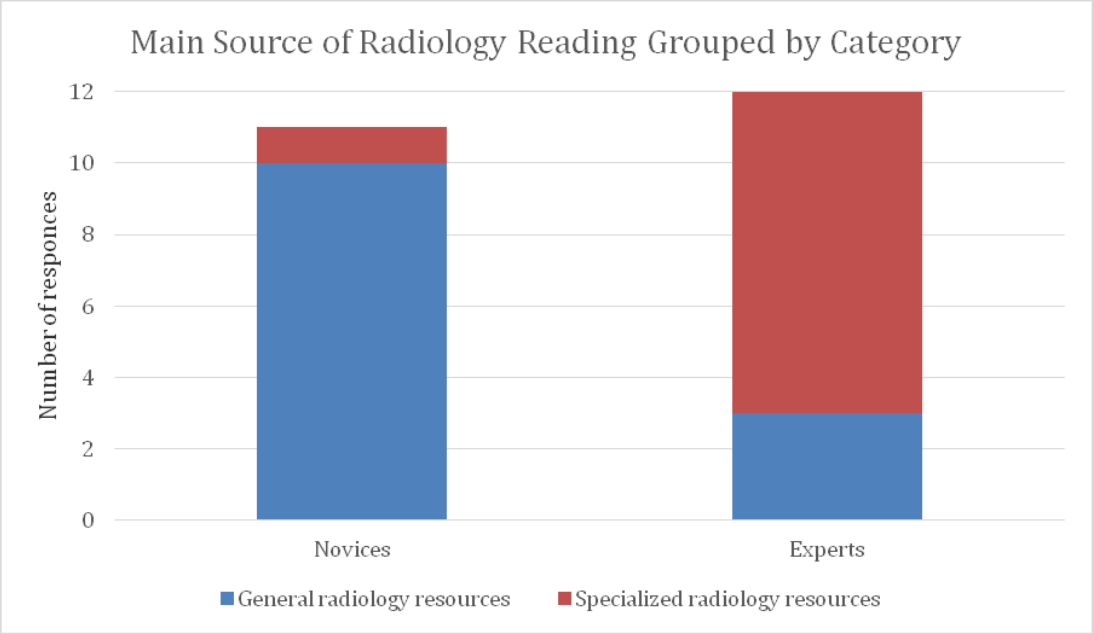


Figure 8. Participants' main source of reading for chest radiology sorted by resource specialization.

Diagnostic Performance Using wAFROC Analysis

wAFROC and Figure of Merit (FOM) analysis demonstrated that the expert group performed better than the novice group in terms of accuracy and confidence levels in reading chest x-rays. FOM was calculated for each participant. As described above, this number reflects the participant's overall performance on all the cases, combining their performance on the frontal and the lateral views. The mean FOM for experts was .59, compared to .52 for novices. The difference between the two was .07. Results of bootstrap analysis, with 2,000 samples, demonstrated that the difference between the two FOM means was significant, p = .038, 95% CI [.005, .136]. This indicates that performance difference is statistically significant and that experts perform better. Figure 9 displays the wAFROC curves for both groups of expertise, representing the difference between them. Figure 10 displays a histogram of the bootstrapped FOM for the novice group, while Figure 11 displays it for the expert group.

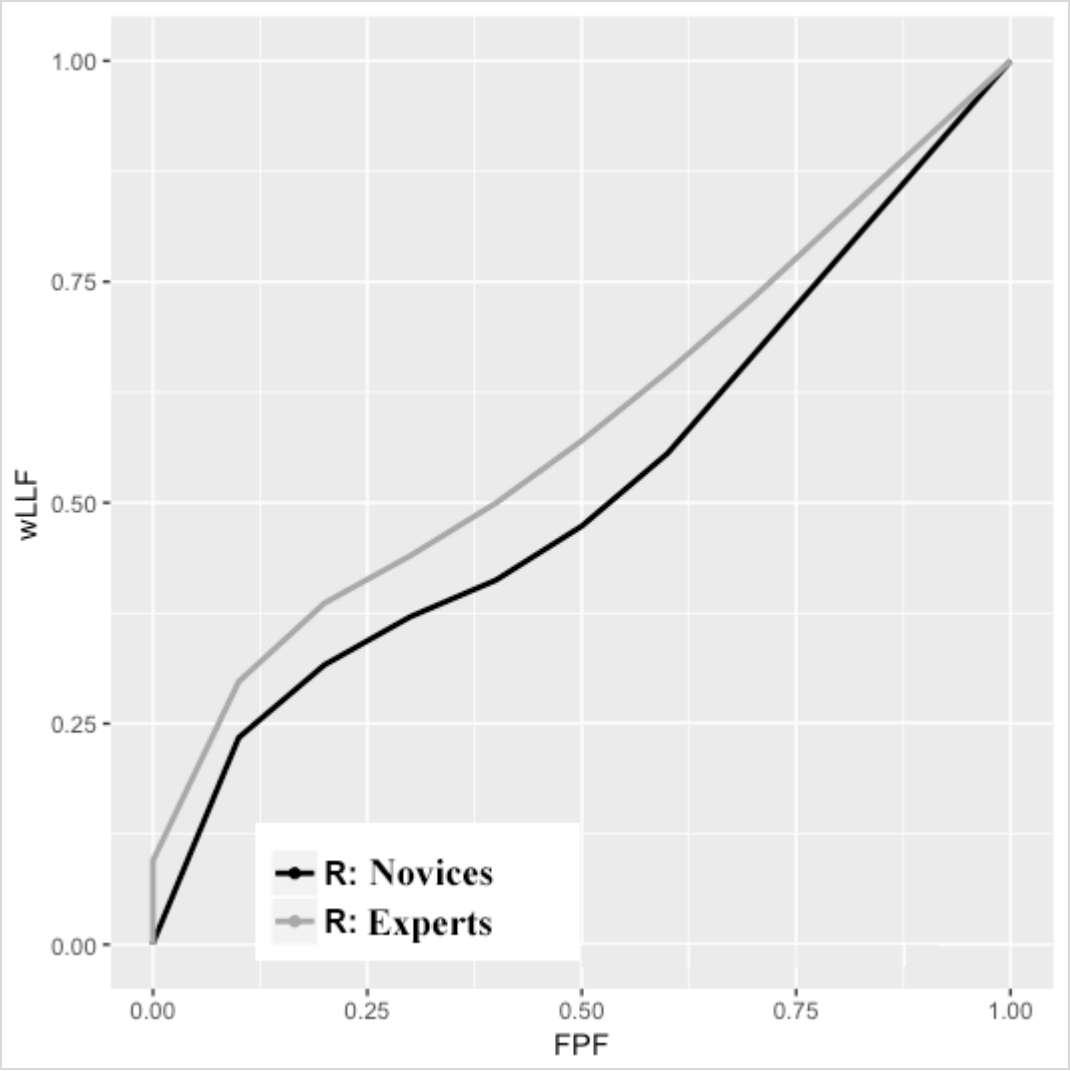


Figure 9. wAFROC curves for both groups of expertise. wLLF stands for Weighted Lesion Localization Fraction. This is derived from the true positive findings participants call. A higher fraction number indicates better performance. FPF stands for False Positive Fraction. This is derived from the false positives participants call. A smaller fraction number indicates better performance.

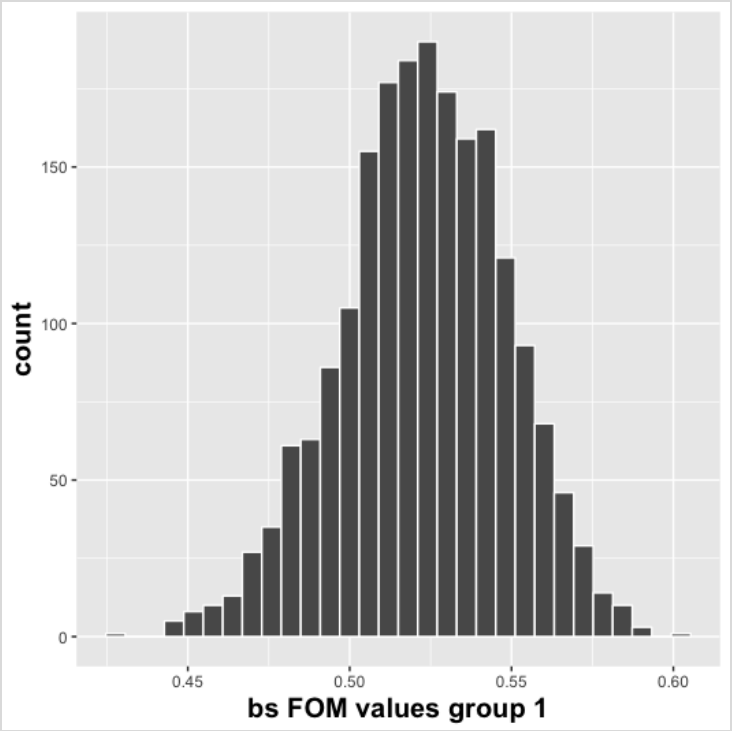


Figure 10. Bootstrapped histogram of the Figures of Merit (FOM) for the novice group.

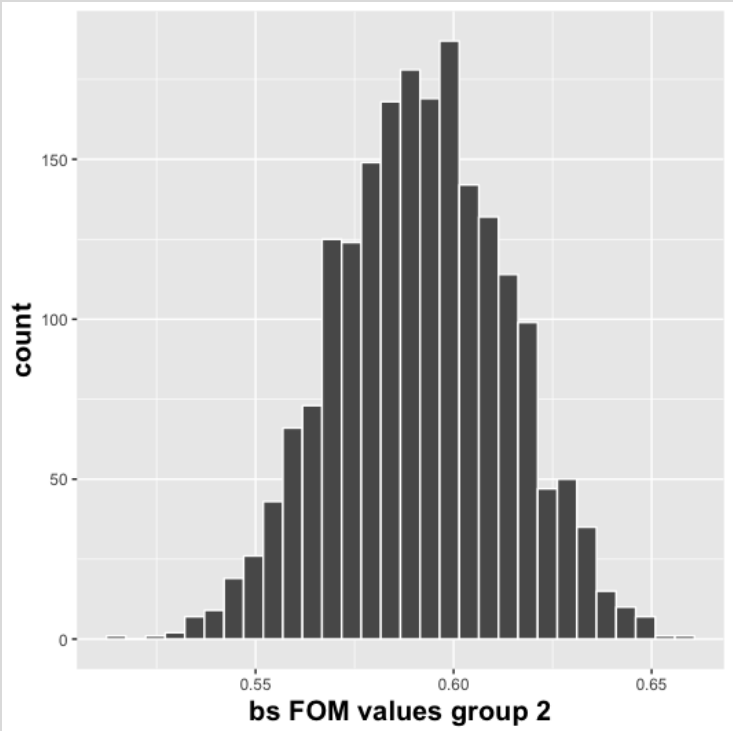


Figure 11. Bootstrapped histogram of the Figures of Merit (FOM) for the expert group.

Eye Tracking Measures and Visual Search Patterns

Novices' mean time to decide whether or not a case was normal, for normal cases was 35373 milliseconds (ms) (*M* = 35373 ms, *SD* = 15796 ms). Experts on the other hand spent 31864 ms to decide on normal cases (*M* = 31864 ms, *SD* = 20181 ms). The difference between the two groups was statistically significant, *t* (481.82) = 2.16, *p* = .031.

Both groups spent less time to decide on abnormal cases compared to normal cases. Novices spent an average of 22792 ms to decide for abnormal cases (*M* = 22792 ms, *SD* = 15387 ms). Experts spent 17737 ms to decide on abnormal cases (*M* = 17734 ms, *SD* = 17064 ms). Again, the difference between the two groups was statistically significant *t* (459) = 3.33, *p* = .001. Figures 12-15 plot the mean time to decide on diagnosis for both groups on normal and abnormal cases.

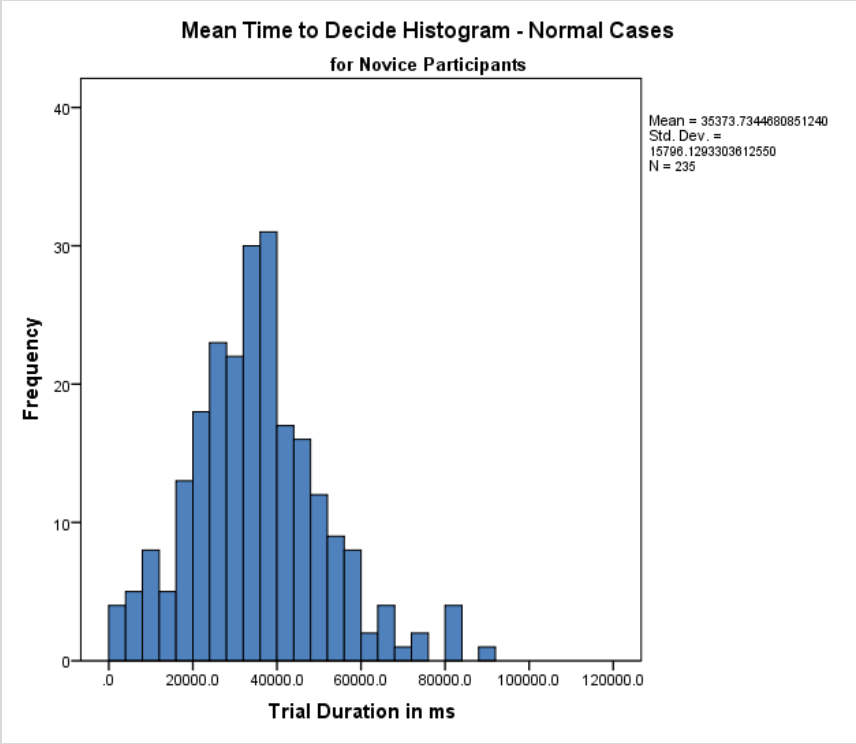


Figure 12. Novices' mean time to decide on normal cases.

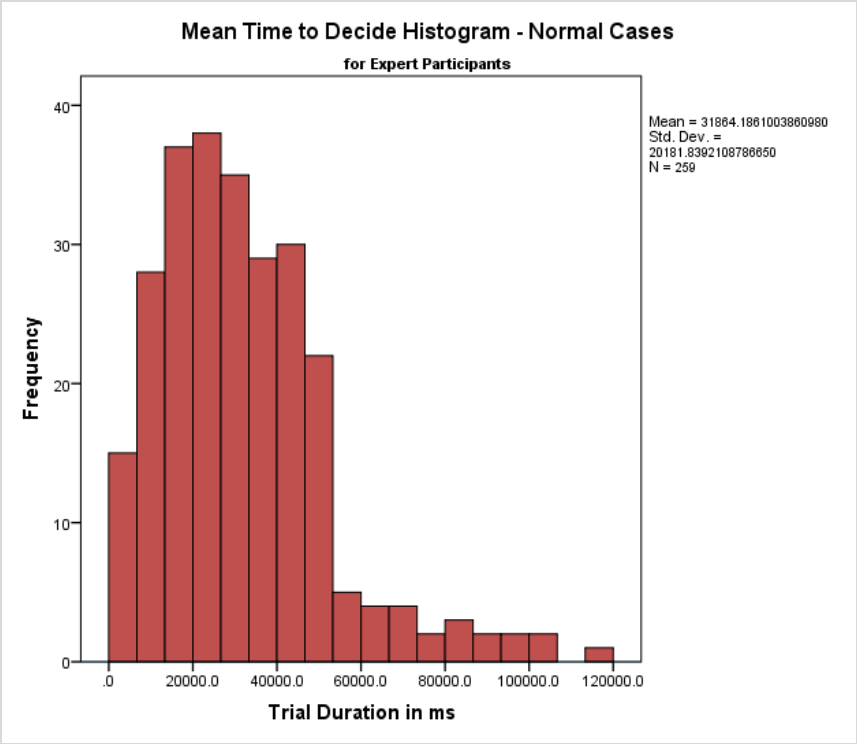


Figure 13. Experts' mean time to decide on normal cases.

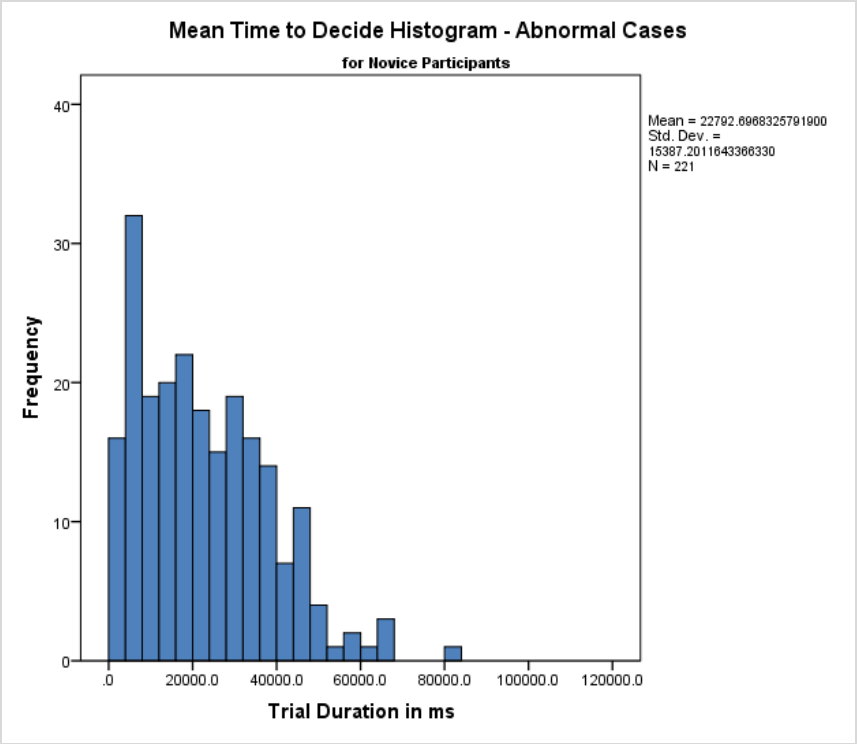


Figure 14. Novices' mean time to decide on abnormal cases. On average, it took novices 22.8 seconds to decide on this group of cases.

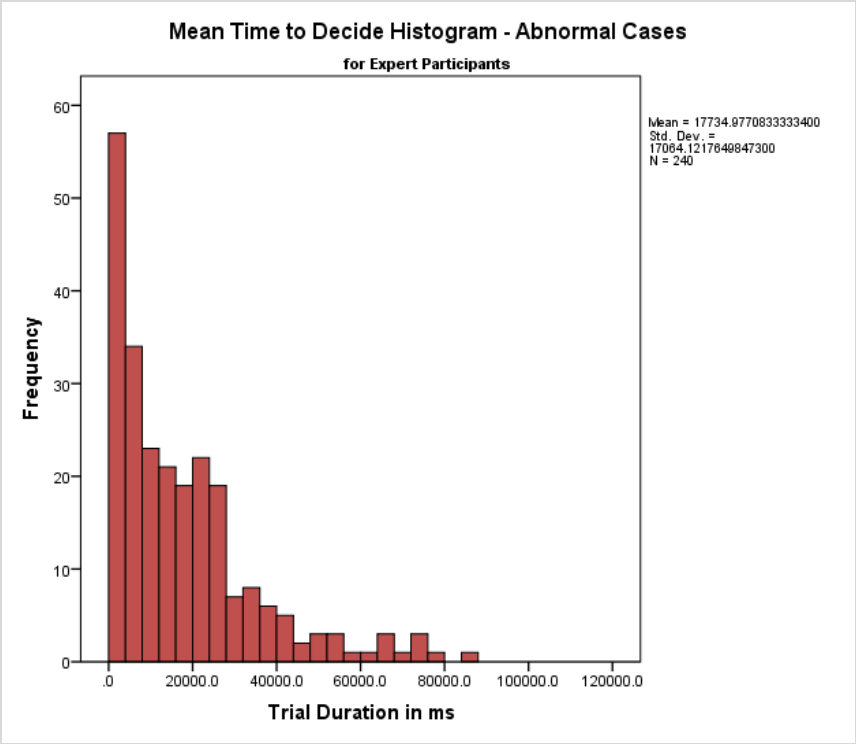


Figure 15. Experts' mean time to decide on abnormal cases. Experts quickly identified the case as abnormal, averaging 17.7 seconds.

A qualitative difference between novices and experts can be appreciated on visual representations of gaze. Scan paths and heat maps were used to provide such qualitative information. An example case of each method is discussed below, as discussing the scan paths and heat maps of 40 cases for each of the 23 participants is not feasible.

Figures 16-18 demonstrate an expert's and a novice's scan paths on an abnormal chest x-ray. The expert identified the abnormality with fewer fixations before deciding on the case and jumped between the PA and lateral views a few times only to confirm a finding seen on the PA. She / he then dismissed the case to give a correct decision. The novice in this example could not identify the abnormality at first glance. It seems that she / he then started to follow a systematic checklist by looking at all regions of the chest x-ray and then mistakenly called the case normal.

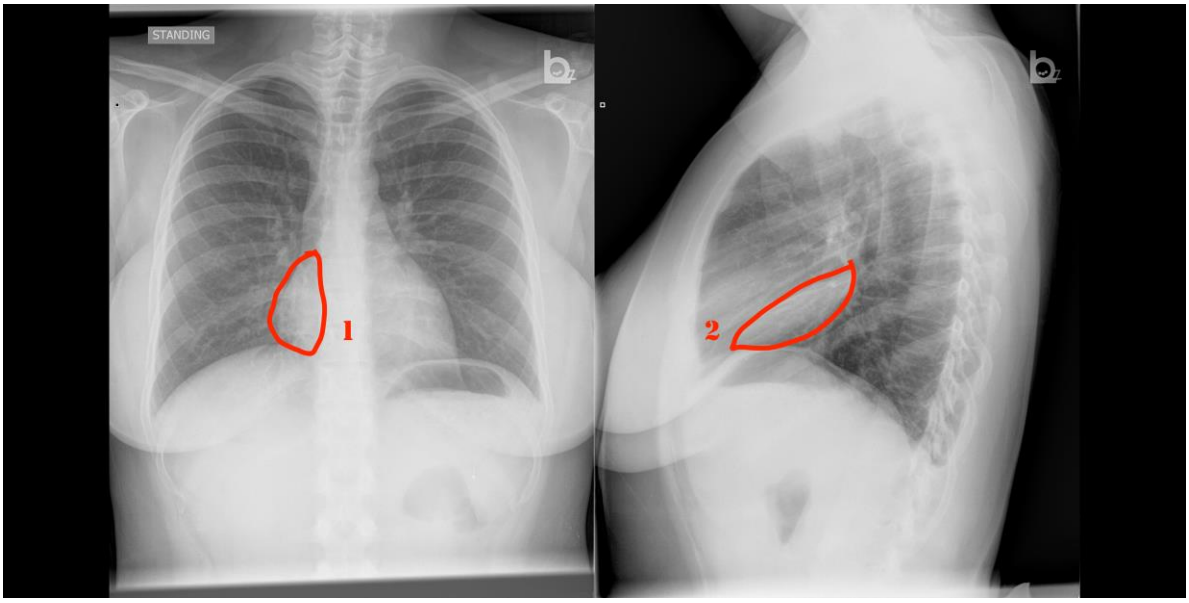


Figure 16. An abnormal chest x-ray with collapse of the right middle lobe. The abnormality is marked in red only in this figure to highlight its location on both views.

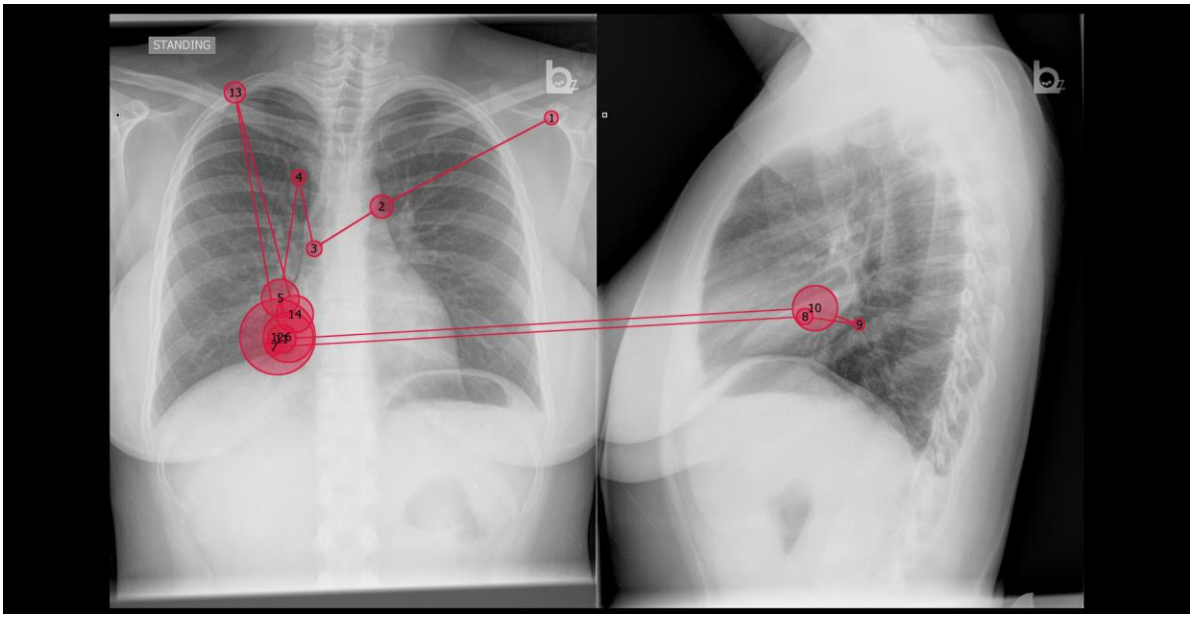


Figure 17. Scan path of an expert participant on the same chest x-ray presented in Fig. 16. The expert identifies the abnormality after a few fixations. She / he then confirms it is a true abnormality by looking at the lateral view (fixations 8, 9, &10) then quickly ends the case and report a correct decision after a total of 14 fixations.

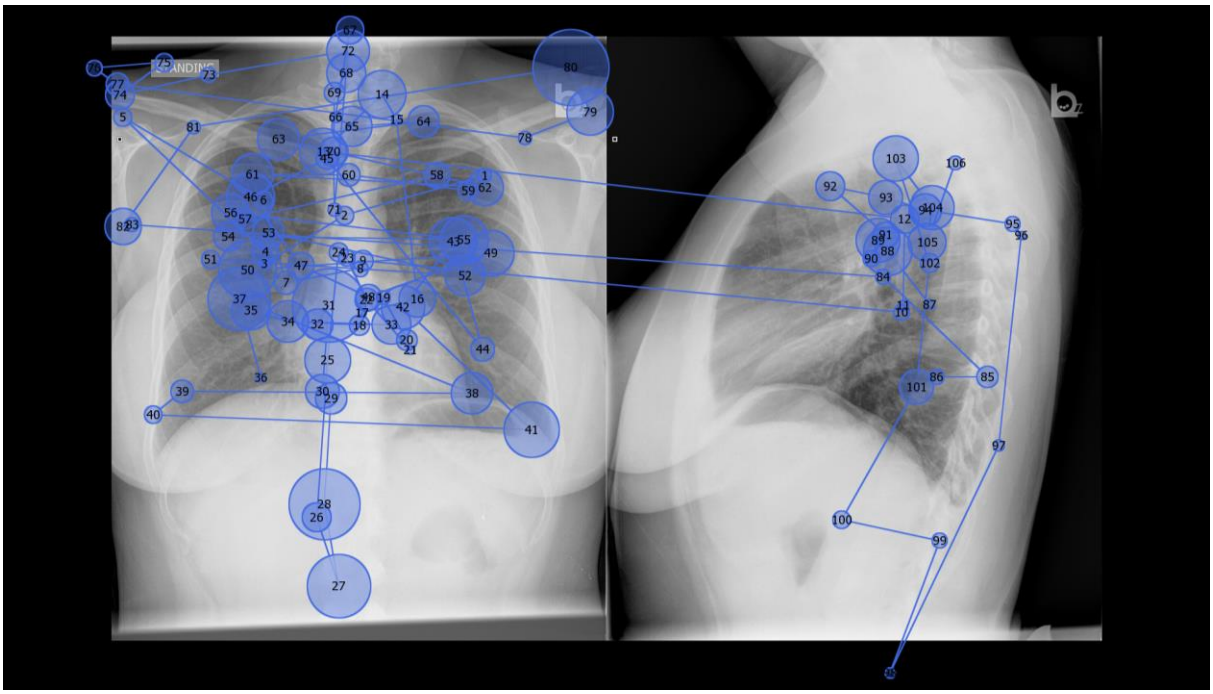


Figure 18. Scan path of a novice participant on the same chest x-ray presented in Fig. 16. The novice in this example could not identify the abnormality at first glance. She / he then starts to follow a checklist by systematically looking at all regions of the chest x-ray. She / he called the case normal as they could not identify the abnormality despite examining it with 106 fixations.

An aggregated heat map of an example case for all participants is shown in Figures 19-21. These figures demonstrate that novices spent more time looking at normal parts of the x-rays compared to experts, whereas experts seemed to “zone-in” to the abnormality in an apparently more direct fashion.

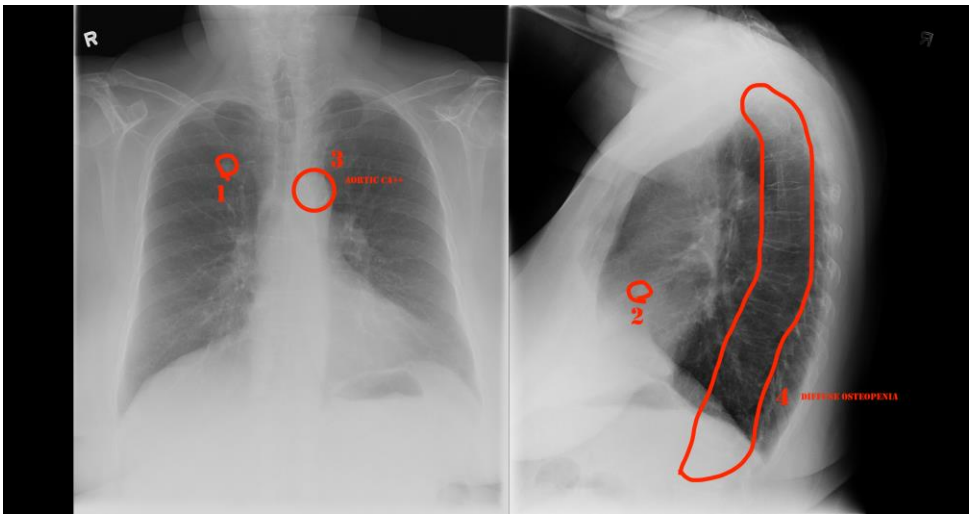


Figure 19. A chest x-ray with multiple abnormalities on the frontal and lateral views. The abnormalities are marked in red to highlight their locations. The following two diagrams (Figs. 20-21) project a heat map on this chest x-ray.

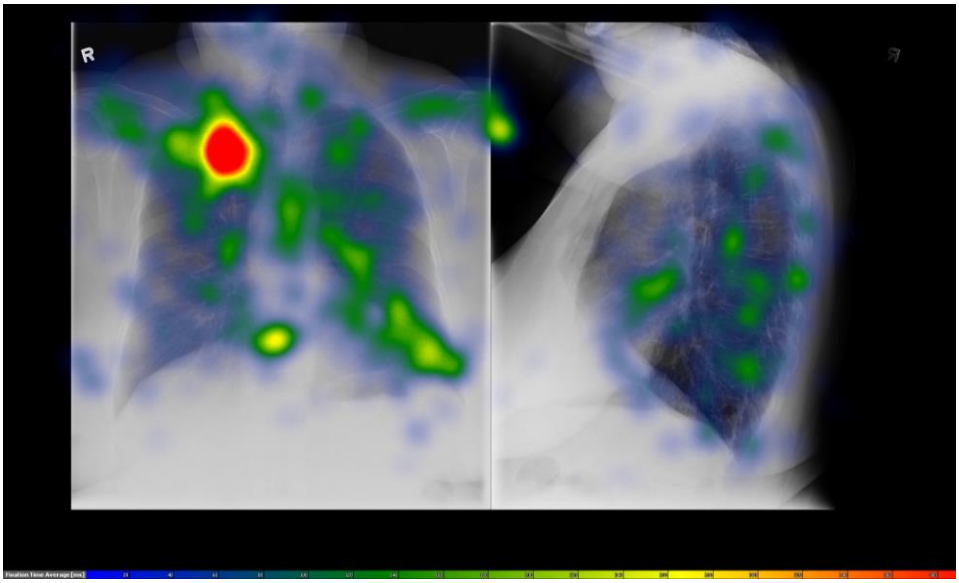


Figure 20. Heat map combining the total fixation times of 12 experts. This is the same case presented in Fig. 19. The scale at the bottom of the image indicates that blue, green, and red correspond to short, intermediate, and long fixation times, respectively. Expert readers averaged shorter fixation times before deciding compared to novices (shown in Fig. 21). A single true abnormality (lung nodule) caught the eye of most expert readers, shown as a red circle.

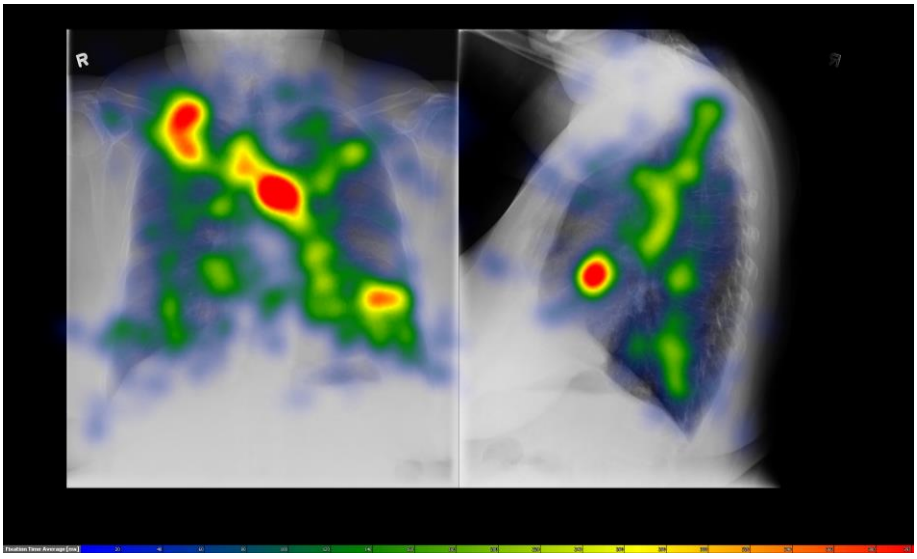


Figure 21. Heat map of the same case displaying the combined total fixation times of the 11 novices. Novice readers spent more time looking, and focused on more than one location before deciding about the case.

False Positives

Our study focuses on the types of false positive mistakes made by each group because a knowledge gap exists in the literature on these types of mistake, as discussed in chapter 2. False positive calls for each group were recorded for each x-ray projection and were analyzed for the most common mistakes. The most common false positive by novices was blunting of the costophrenic (CP) angles; n = 37 in the frontal view, n = 30 in the lateral view. This was followed by a lung opacity; n = 35 in the frontal view, n = 15 in the lateral view. The third false positive was increased lung markings; n = 34 in the frontal view, n = 11 in the lateral view. Experts differed from novices in the type of false positive errors. The most commonly called false positive by the expert group was a lung opacity; n = 65 in the frontal view, n = 35 in the lateral view. This error is followed by reticular opacities; n = 24 in the frontal view, n = 6 in the lateral view. The third was a mediastinal mass; n = 15 in the frontal view, n = 7 in the lateral view. Figures 22-23 display the most common false positive mistakes by each group on each projection.

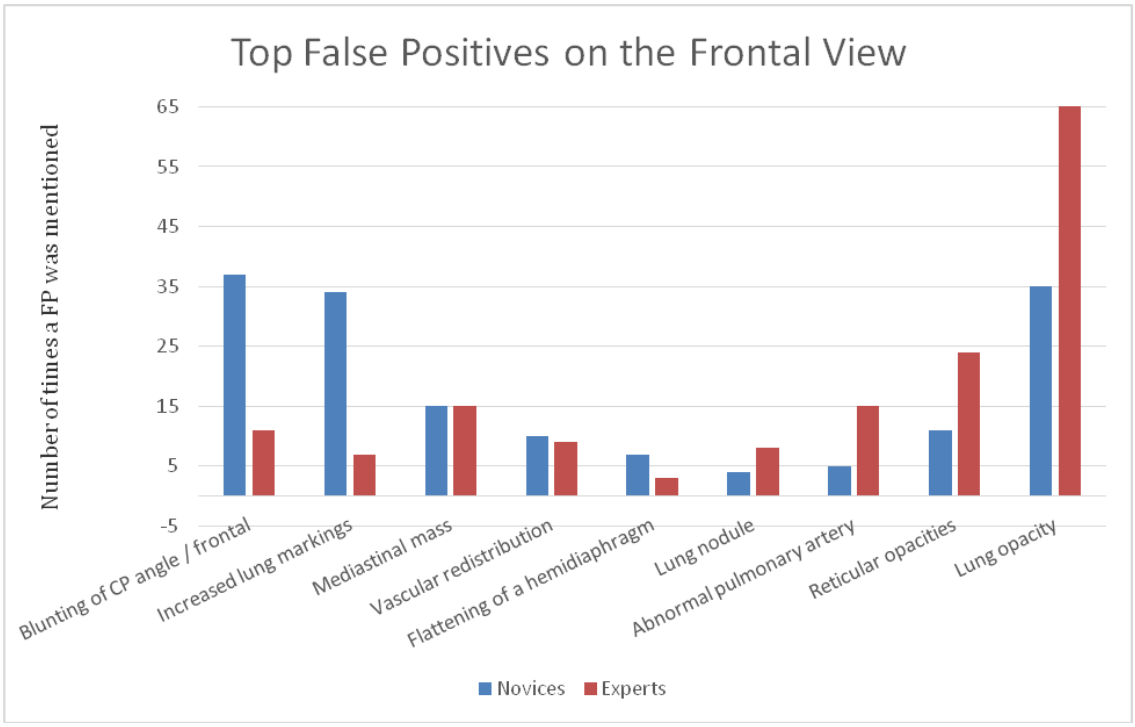
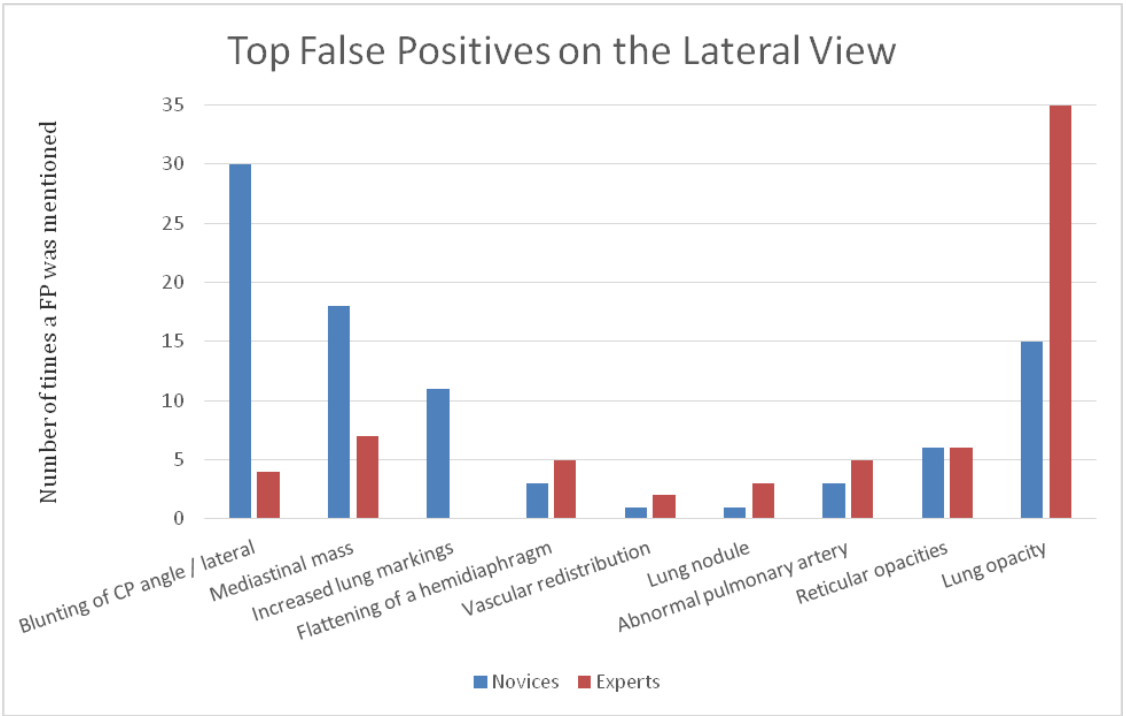


Figure 22. The most common false positives called on the PA view by each group of expertise. The mistakes are arranged so that novices' FP are on the left and experts' FP are on the right



.

Figure 23. The most common false positives called on the lateral view by each group of expertise. The mistakes are arranged so that novices' FP are on the left and experts' FP are on the right.

Similar to scan paths and heat maps, focus maps are another type of qualitative visual representation of gaze. They are generated by altering the transparency of the image so that areas that received more gaze attention are more transparent. Figure 24 and Figure 25 are example focus maps of participants' gaze on a normal chest x-ray. The figures demonstrate that novices have longer gaze on areas commonly called false positive. Collectively, novices paid more attention to the costophrenic angles, an area that they commonly falsely call as abnormal. They also focused more on the mediastinum, especially on the lateral view, another more commonly called false positive by this group.

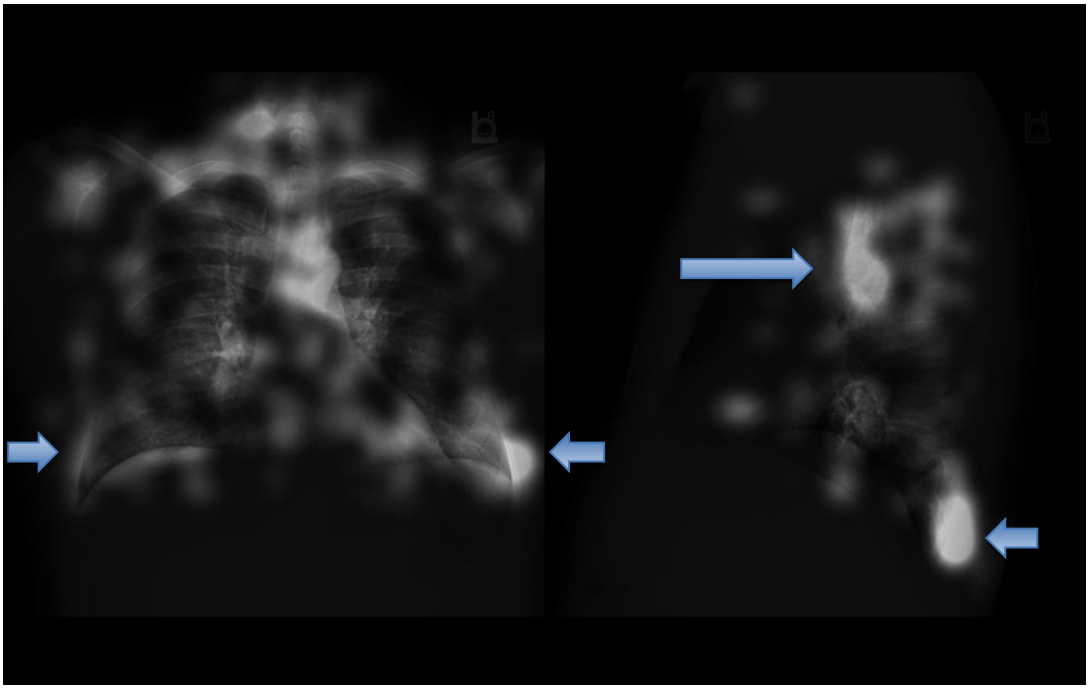


Figure 24. A focus map on a normal chest x-ray highlighting the combined areas of focus of the 11 novices. This group focused much on the costophrenic angles (short arrows) and the mediastinum (long arrow). These areas had very high false positive reports by novices. Additionally, less area is covered overall in comparison to experts (Fig. 25).

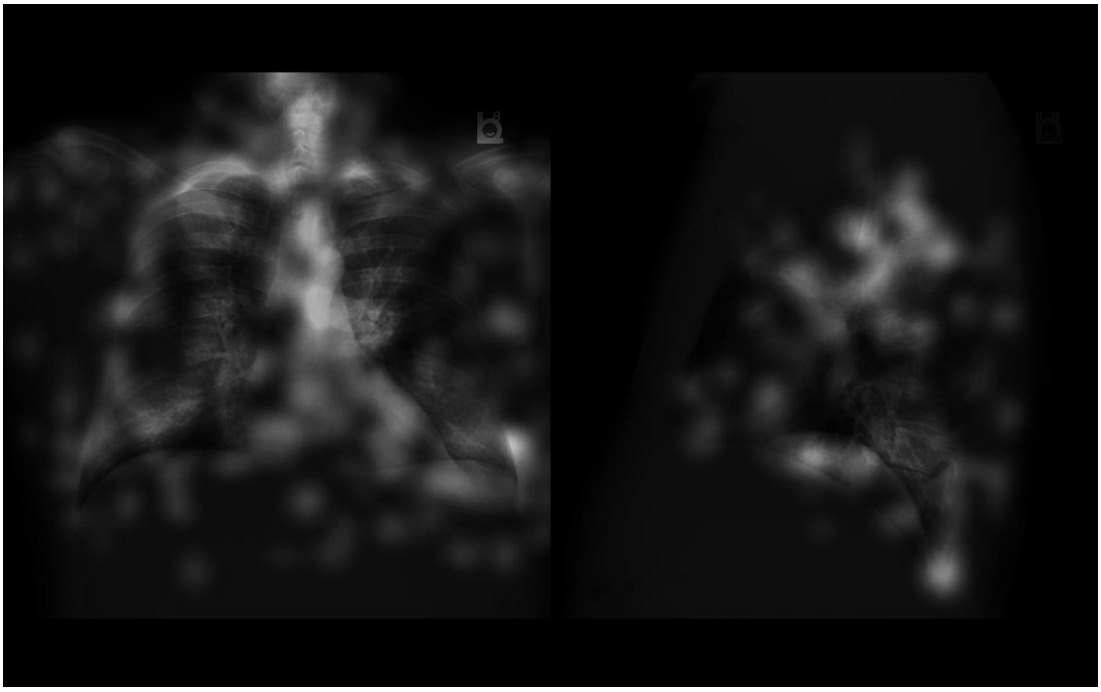


Figure 25. A focus map on a normal chest x-ray highlighting the combined areas of focus of the 12 experts. This group covered a larger overall area in comparison to novices and focused less on the costophrenic angles and the mediastinum. The lateral view is covered more homogenously by experts than by novices.

Special False Positive Cases

There were interesting observations during the experiment that prompted a focused post-hoc analysis on areas commonly discussed in the literature. For example, many participants falsely called a normal appearing anatomical area known as the “retrocardiac clear space” abnormal. This and similar observations incited a review of the mistakes made on important normal anatomical/radiological areas on the frontal and lateral films. Figure 26 demonstrates the differences between experts and novices on these special areas.

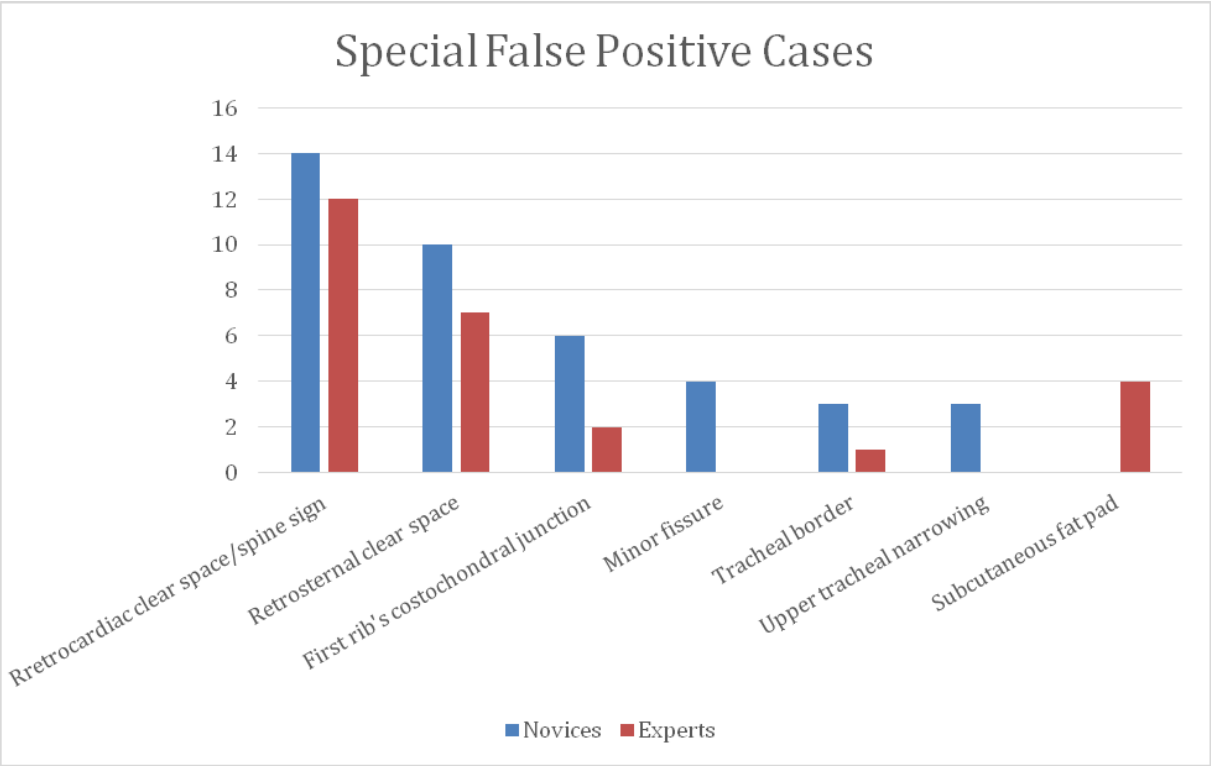


Figure 26. Special false positive cases. Participants' clicks on specific normal anatomical and radiological landmarks, calling them abnormal. The fat pad was called abnormal four times in a single case by a 4th-year resident, potentially making it an outlier.

One participant demonstrated an active learning process during the experiment. For example, one 2nd-year resident (novice) mistakenly called an "upper tracheal narrowing" on a normal case. The following case was also normal and the participant again, hesitantly and after a longer period of time, called an "upper tracheal narrowing." When the third case showed up, the participant immediately looked at the upper trachea, laughed, and said, "ok, that's how it should look." She didn't call it abnormal that time.

Think Aloud Analysis

The think aloud analysis yielded information in several areas: (a) the accuracy of diagnosis, (b) the number of anatomical structures enumerated by each group, (c) attention to the lateral view, (d) the thought process that lead to the diagnosis, and (e) the thought process after making a diagnosis.

As mentioned, there were two think aloud cases; one was normal while the other one was abnormal. Starting with the normal case, all participants called it normal, with no false positives. The mean number of anatomical structures mentioned by novices in the normal case was 18.5 structures (M = 18.5 structures, SD = 8.2 structures). Experts mentioned a slightly smaller number of structures (M = 17.6 structures, SD = 10 structures). The difference between the groups was not statistically significant, t (21) = .25, p = .804.

Further analysis of the anatomical structures based on each view of the chest x-ray showed that novices on average mentioned 15.7 anatomical structures on the PA view (M = 15.7 structures, SD = 7.5 structures). Experts mentioned a slightly smaller number of structures on that view (M = 12.7 structures, SD = 7.3 structures). Again, the difference between the groups was not statistically significant, t (21) = 1.04, p = .308.

For the lateral view, novices on average mentioned five anatomical structures (M = 5 structures, SD = 3.7 structures). Experts mentioned a slightly higher number of structures (M = 5.8 structures, SD = 3.8 structures); however, the difference was not statistically significant, t (21) = -.53, p = .6.

When counting the number of times the lateral view was mentioned during the think aloud, novices mention it 2.2 times on average (M = 2.2 times, SD = 1.7 times). Experts mentioned the lateral a slightly higher number of times (M = 2.6 times, SD = 1.8 times). Again, the difference between the groups was not statistically significant, t (21) = -.54, p = .6.

For the second think aloud case (the abnormal case), the participants' final diagnoses were recorded as a first step in the analysis. This showed that one novice (9.1% of novices) and two experts (16.7% of experts) called the case normal. A main feature of that case was the presence of a previous mastectomy (surgical removal of breast tissue). Six novices (54.5% of novices) and seven experts (58.3% of experts) found the mastectomy. Absence of a normal

anatomical structure can sometimes be harder than identifying an abnormal one, so the thought process behind identification of the mastectomy was studied. There were three ways participants identified that abnormality. Some identified an overall lung density discrepancy between the right and left lung, which led them to search for a reason behind that finding, finally identifying a missing breast shadow. The second pathway was identifying a nodule in the right lower lung where the normal breast is present, then suspecting that this nodule actually represented the normal nipple shadow, leading them to an active search of the other breast's nipple shadow for comparison. Finally, they identified the previous left sided mastectomy. In the last pattern, participants mentioned a previous mastectomy by basically calling it directly: "there is a left side mastectomy." It was not possible to statistically test whether a difference exists between novices and experts in terms of the paths they follow because the number of participants who identified the mastectomy was low. Table 1 shows the distribution of the three pathways between the two groups.

Table 1

The Thought Process Behind Identifying a Mastectomy

Thought process initial finding	Novices	Experts
Identifying a lung density discrepancy	2	3
Identifying the contralateral nipple shadow	1	2
Straight calling out a mastectomy	3	2

Note. The initial finding led to further analysis of the image, leading finally to identification of a missing anatomical structure.

One interesting and important feature of this case is that all the findings relate to the previous mastectomy. In fact, the patient had a mastectomy due to breast cancer and that surgery included an axillary lymph node dissection. This shows on the PA and lateral chest x-rays as

metallic clips in the patient's axillary area. She also has a collapse of the right middle lobe of the lung that is due to radiotherapy treatment of her cancer. Due to the relationship between these radiological findings, it was important to study any emerging patterns that distinguished experts from novices in their approach to such a complex case.

This exploration showed that for some participants, the presence of a previous mastectomy triggered a search for other findings. Only one novice (9.15% of novices) stated that they will look for specific findings because of the previous mastectomy. On the other hand, five experts (41.7% of experts) mentioned that they will proceed with such a search due to the presence of the previous mastectomy. In this case, finding a previous mastectomy should trigger a search for eight items in the frontal view and seven on the lateral. (These are shown in Table 2.) This list of items is theoretical and based on suspected abnormalities that could be present as a consequence of breast cancer and previous mastectomy. Out of this list, the case actually has three findings on the frontal and two on the lateral. On average, novices mentioned that they looked for .6 out of eight items on the frontal (M = .6 items, SD = 1.8 items) and .3 out of seven items on the lateral (M = .3 items, SD = .9 items). They found .7 items on average on the frontal (M = .7 items, SD = .9 items), and .5 items on the lateral (M = .5 items, SD = .5 items). Experts on the other hand mentioned that they looked, on average, for 1.6 items on the frontal (M = 1.6 items, SD = 1.3 items), and .3 items on the lateral (M = .3 items, SD = .7 items). They found .7 items on the frontal, (M = .7 items, SD = .7 items) and .5 items on the lateral (M = .5 items, SD = .5 items). There was no statistically significant difference between the groups regarding the total number of items searched for in the frontal view, t (21) = -1.45, p = .16, nor in the lateral, t (21) = -.19, p = .86. There was also no statistically significant difference with regards to the number of items found in the frontal t (21) = .19, p = .86, nor the number of items found in the lateral t (21) = -.21, p = .84.

Table 2

Items to Be Searched for in the Mastectomy Case

Items to search for in the mastectomy case	Frontal	Lateral
The presence of axillary lymph nodes	✓	✓
Evidence of radiation pneumonitis	✓	✓
Cause of RML collapse	✓	✓
Lung nodules	✓	✓
Pleural effusions	✓	✓
Bone lesions	✓	✓
Hilar adenopathy	✓	✓
Supraclavicular adenopathy	✓	

Note. Finding a mastectomy in this case should trigger an active search for these items.

CHAPTER 5: DISCUSSION

The purpose of the study was to add to the literature on radiology expertise by investigating a research gap. This gap involves examining how experts and novices utilize the different views of the chest x-ray in making diagnostic decisions. In particular, this study examines expert–novice differences in reading the frontal and lateral chest radiographs using reliable and well-established research methodologies: eye tracking and think aloud. We review these expert–novice differences below.

As expected, the expert group we studied reported a greater number of years of experience in radiology and a higher number of chest rotations when compared to the novice group. Experience is an important cornerstone in the path to expertise. Exposure to a high number of cases builds a subconscious repository and a mental schema of normal and abnormal features so that decisions become well informed (Manning, 2010). Greater experience levels provide more opportunities for feedback on decisions and fine-tuning them based on real life experience. Experience also provides an opportunity to apply theoretical knowledge into everyday practice (Manning et al., 2006; Nodine et al., 1999).

Interestingly, none of the novices reported an intention to specialize in cardiothoracic radiology. This finding might represent lack of interest in this specialty and may have bearing on the performance of the novice group. This point, however, could be seen as an advantage to the study as it might widen the performance gap between the groups and shine more light on performance differences between experts and novices. It is important to note, however, that despite the novices' low interest in pursuing cardiothoracic radiology as a specialty, they are still required to master a basic level of that specialty, and hence were an appropriate group for the aim of this study.

When comfort levels were examined, experts turned out to be significantly more comfortable in reading PA chest x-rays and chest CT scans. This was an expected finding as

experts are more comfortable handling even complex tasks in their domain of expertise. However, when it came to the lateral chest x-ray, no group differences were reported in comfort level. There could be multiple reasons behind this unexpected finding, related either to lower experts' comfort level or higher novices' comfort level. Both situations can explain the relatively similar mean and standard deviation of both groups (novices reported a mean comfort level of 3.3 and standard deviation of .65, while experts reported a mean of 3.8 and a standard deviation of .97). Looking at Figure 6 once more, we see that the lateral film is the only place where experts reported being "uncomfortable." At the same time, most novices reported a "neutral" response. This finding may be explained by the Dunning–Kruger effect that explains how novices think highly of their abilities when they are starting to learn about something, not knowing the difficulties and details that are there. As they progress, they start to recognize their weaknesses and their self-confidence becomes lower. Eventually, with more training, they regain some of that confidence (Kruger & Dunning, 1999). In our case, novices might initially be unaware of how complex the lateral film is and report intermediate or high comfort levels. Later on, as training progresses and they attain intermediate expertise, they may report lower comfort levels after understanding the difficulties of the lateral film.

Overall, the lateral chest x-ray had the highest number of "uncomfortable" ratings. This finding supports the argument that these radiological studies are difficult and emphasizes the need for additional teaching and training on how to interpret them. It also concurs with the necessity of additional research about the nature of expertise in this important imaging study.

Participants reported 12 resources as their main sources of reading for chest radiology. Such information is quite important to any learner seeking reliable and trusted reading material. This list stands as a reference for trainees and junior staff, providing an overview of important and high-yield reading resources for chest radiology. To our knowledge, not many articles

describe top reading resources for trainees in the field of radiology, and future research needs to focus on answering this particular question for the whole spectrum of radiology subspecialties.

Our results show that experts and novices read from two completely different types of resources, one being specialized and the second being general, respectively. As expected, experts seek deeper knowledge and state of the art information found in specialized cardiothoracic radiology resources and journal articles. Novices, on the other hand, search for basic but important information to start with, which is found in general radiology resources; they are still acquiring basic knowledge and need to read at a level that supports their current understanding, while experts have a lot of knowledge already and seek out more specialized resources. One reason behind this preference is that residency examinations and assessments in early training require a level of knowledge that usually is provided sufficiently by general radiology resources. Another theoretical explanation for novices seeking out general resources can be described by cognitive load theory (Sweller, 1994). This theory discusses the limitations of memory and advises decreasing the cognitive load on working memory to facilitate learning (Young, Van Merrienboer, Durning, & Ten Cate, 2014). General radiology resources have shallower information and therefore lower cognitive load on novice trainees compared to experts. Novices therefore prefer them to specialized resources. Another reason behind the different resources

Experts in our study were faster and more accurate at making diagnostic decisions than novices—they solved the same cases in significantly less time. The quicker expert response is a common dimension of expertise (Chi & Glaser, 1988; Kundel & Nodine, 1975). In fact, a previous study by Kundel and Nodine (1975) showed that a surprising accuracy of 70% for a true positive lesion localization was achieved by expert radiologists even when chest x-rays were displayed for only 0.2 seconds.

One explanation for the differences in time spent on cases is that experts have more experience; novices on the other hand are following their training guidelines and conduct a

systematic search pattern, which may take more time. Radiology educators have long advised trainees to follow systematic approaches that include checklists to minimize misses, especially when no abnormality is identified on the image (Berbaum, Franken, Caldwell, & Schartz, 2010; Kondo & Swerdlow, 2013; Subramaniam, Sherriff, Holmes, Chan, & Shadbolt, 2006). This systematic search process expectedly increases search time, especially when images are perceived as normal, but in theory would provide "full coverage" of the diagnostic image, hoping it would help to identify subtle abnormalities.

Experts generally make fewer errors than novices. However, the types of errors that are made by both groups provide insights into how to improve training and enhance patient care. Knowing the mistakes novices make more than experts is the first step in developing training methods to rectify those weak spots in novices' performance. As mentioned earlier, not all mistakes are the same. There are false negative mistakes, which occur when a finding is missed and an abnormal case is called normal. The danger of this type lies mostly in missing important diagnosis and delaying treatment. Then there are also false positive mistakes. These occur when a wrong diagnosis is given to a normal x-ray, which can lead to unwarranted worry, investigations, and unnecessary cost and treatment. Due to significant consequences of false negatives in delaying diagnosis, they have been studied extensively in the literature. Expertise in false positives on the other hand is an area of research that is just being tapped, with very little if any published literature. An analysis of false positive errors, reporting abnormalities when there are none, is an important part of this research. In general, novices had lower performance scores, as measured by FOM, than experts in our study. This implies that they made more errors and were more confident about them. This overconfidence in incorrect findings would be dangerous if residents were not supervised and that is one reason cases are always supervised until residents graduate from their training. Novices also made fewer true positive decisions and were less

confident about them compared to experts. This indicates they miss more radiological findings in comparison to experts, and even when they did identify an abnormality they were still not highly confident about it. The most commonly made false positive mistake by novices was "blunting of the costophrenic angle." This finding is interesting, and could be explained by some lack of expertise reading the lateral view amongst novices. True costophrenic angle blunting most commonly occurs when fluid accumulates around the lung, in which case the blunting is evident on *both* the frontal and lateral views. The costophrenic angle can sometimes appear falsely blunted on the frontal view, even in the absence of fluid, in which case checking the lateral view and finding the angle sharp and clear will allow the radiologist to dismiss the apparent blunting on the frontal view. A certain degree of expertise is needed to remember to do this, and to do it correctly. Indeed, it was apparent from the recordings of some novices' mistakes that they did not confirm their opinion by corroborating with the lateral view.

In another instance, the investigator reviewed a novice's mistakes with her after she finished her experiment. It was apparent that she overcalled blunting of the CP angle, and upon discussion, she mentioned that she was following strict "by the book" criteria for that radiologic finding. Any CP angle that is not pointed enough to "prick you if you virtually touched it" is abnormal. Some cases she overcalled were completely normal apart from a questionably blunt CP angle. Following that criteria, it is understandable why she called CP blunting in such cases; however, such minimal blunting seen alone is almost never called in everyday practice. None of the expert panel members did call that finding on those cases. This emphasized the importance of taking the case as a whole rather than fragmenting each finding separately. This holistic approach by experts is also seen in the abnormal think aloud case and will be discussed shortly.

An unexpected finding in our study was that experts tended to identify lung opacities when they were absent. This finding is interesting and warrants some thought. A suggested

reason behind this result is that a lung opacity is one of the most commonly seen abnormalities and could represent a wide variety of underlying pathologies. It is therefore an "everyday finding" and would be called more by practicing radiologists (experts) who view a lot of cases on an everyday basis and are used to commonly seen pathologies. Additionally, lung opacities and nodules are common presentations of lung cancers, an important "do not miss" diagnosis. At the same time, they are one of the most commonly missed diagnoses (Berlin, 1986; Hamer, Morlock, Foley, & Ros, 1987; Spring & Tennenhouse, 1986). Experts could be more aware of this fact and thus more cautious when seeing something suspicious of a lung nodule. Novices on the other hand are connected more to textbooks and theoretical manuscripts that discuss various diseases, despite some of these diseases being less common. If we consider these background differences it can help explain why more experts called false positive lung opacities and lung nodules, as these are more commonly encountered in everyday practice.

As expected, novices made more false positive mistakes on normal anatomical structures in both the frontal and the lateral views, including the upper trachea and its border, the minor fissure, and the tip of the first rib. One explanation for this finding is that novices are less exposed to normal cases. These trainees have only recently entered the world of diagnostic radiology and their understanding of diseases is mostly based on abnormal cases shown in their reading resources. They have much less experience with real life cases, many of which are normal. Their recent studying and beginner's enthusiasm may prime them to diagnose chest x-rays with what they have just read about, or at least some sort of abnormality. It seems as if they are thinking: "most radiologic studies require some sort of diagnosis and cannot be just normal." Add to that the fact that trainees are never examined on normal cases; they are usually assessed on abnormal cases and are expected to find abnormalities.

Our results suggest that the lack of exposure to normal cases could be overlooked in radiology training. Trainees do get better with time, as seen in the case of experts, but this process of familiarity with the normal chest x-ray could perhaps be hastened by deliberate practice and exposure to more of them during training and assessment. Everyday reporting partly fulfills this requirement, but teaching rounds, lectures, and exams usually do not.

Interestingly enough, all participants correctly called the normal think aloud case normal. Despite many novices and experts calling false positives in other normal cases, their performance on this case appears better than when they were not asked to verbalize their thought process. Performance improvement with verbal think aloud is a known phenomenon and has been described in the literature. Think alouds might enhance performance by changing the sequence of thinking during the problem solving process (Ericsson, 2006b). Another possible cause of better performance on this case is that it was presented at the end of the experiment and participants would have seen many normal and abnormal cases prior to seeing it, therefore improving performance towards the end (practice effect bias). The simplicity of the case might have been a factor as well, being a case with little questionable findings on it. Regardless of the reason of their better performance, this case turned out to be a good opportunity to understand the thought process while reading a normal case for all of our participants.

The close equivalence of the number of anatomical structures mentioned by each group is quite interesting because one would expect experts to enumerate more structures when viewing an x-ray compared to novices; however, this was not the case in our study. Our results show that both groups have a similar number of items in their mental checklists when viewing a normal chest x-ray. One possible explanation of this similarity is the proximity of 2md- and 3rd-year

residents in terms of expertise. Based on clinical norms; 3rd-year residents are considered seniors and take greater responsibility compared to 2nd-year residents.

Regardless of this issue, one would think the similar number of anatomical structures mentioned by each group indicated their searches were equally thorough. Why then did novices make more mistakes in our experiment as a whole? There could be more than one answer for this question. Although novices and experts may have a similar number of items on their mental checklists, the nature of these items may be different; novices may be exploring less important, general items while experts may be focusing on important items. Looking at expertise in radiology literature, we see that errors stem mainly from: (a) failure to observe an abnormality; (b) failure to recognizing an abnormality, despite having observed it; (c) wrong interpretation of a radiological finding; or (d) failure to communicate an abnormality properly (Pinto & Brunese, 2010). It is possible that novices in our study often succeeded in the observation part but fell short in the recognition and interpretation parts. In general, this explanation is in coherence with what the literature reports. A recent study questioned the benefit of systematically reviewing (e.g., using anatomical checklists) a radiologic study, as their results did not demonstrate performance improvement with such techniques (Kok et al., 2016). Again, their finding supports the importance of what to look for "in each item in the list" in addition to what to look for "in the image in general." The grain size should be smaller and attention to details is important, not just going through a list.

Looking at the second think aloud case, it was surprising to see that two experts called it normal when it was abnormal, despite the multiple findings on both the frontal and the lateral views. These misses could be because participants were fatigued at the end of the experiment. Research has indeed shown that fatigue is an important contributing factor to diagnostic error in

radiology (Lee et al., 2013). They may also have felt rushed to finish the experiment and continue on with their busy schedules, providing a "premature closure" (Lee et al., 2013) to their interpretation of the case. Another possibility is that those misses could be part of the error rate reported in the literature (Fitzgerald, 2001; Lee et al., 2013).The uninterrupted think aloud protocol we used did not allow for root analysis discussion to understand the exact reason for missing this significant finding. However, an important thing to bear in mind in this particular case is that the main finding is actually the absence of a normal anatomical structure—the left breast. William J. Tuddenham, an early pioneer in the field of medical image perception, published about reader error long ago. In 1962 he reported that gross and obvious findings were often overlooked. As he describes, "two of our three readers failed to note the complete disarticulation of the shoulder girdle… and the most commonly missed finding in the study was the amputation of the female breast" (Tuddenham, 1962, p. 701). Identifying common pathways that lead participants to discover such a difficult finding (Table 1) is an important step in understanding the thought process that leads to identifying a previous mastectomy. Understanding such pathways used by experts can help expand novices' search patterns and mental checklists. Our study is considered a pioneer in the radiology expertise literature in the sense that it uncovers expert thinking processes and common pathways that lead to identifying a difficult radiological feature.

Some experts not only outperformed novices in terms of finding the mastectomy but also went beyond that; they searched for findings *associated* with it. Their think aloud showed that they systematically looked for other likely findings (axillary clips, lung collapse, etc.), which reflects their superior domain knowledge and deeper understanding of the mechanisms and consequences of diseases, a finding that is documented in the expertise literature (Manning, 2010; Nodine & Mello-Thoms, 2010).

Limitations

As with other scientific research, our study encountered some limitations that should be considered when viewing the results. These limitations are not major, in our opinion, and do not notably affect the results.

It would have been ideal to have participants from the 5th year of residency. One would expect such participants to show superior performance due to being in their final year, and preparing intensely for their examinations. However, no residents in that year volunteered to participate, probably for those same reasons.

Our eye tracking device showed lower tracking abilities for participants with eye glasses. This decreased the accuracy of fine and detailed eye tracking metrics such as *target* dwell time. However, the main purpose of the eye tracking in this study was not dependent on such fine metrics, but rather on more reliable metrics such as *total* dwell time. This limitation might affect future post hoc analysis using the same data set and not our main study outcomes. Upon discussion with the providing company, the technical support team explained that this is an inherent limitation of the model in hand and that later models show better performance with participants wearing eye glasses.

The study was conducted in four different rooms in two hospitals. It was quite difficult therefore to standardize the room conditions, including light and noise levels. Similar conditions were sought but it was not always achieved. Lighting can affect performance and ideally all participants should be examined in similar conditions. Unfortunately, that was not always the case. No participant, however, complained about too much or too little light and overall the effect of different ambient lighting conditions is not expected to be significant, as shown in previous research (Pollard et al., 2012).

One limitation of the study relates to the type of wAFROC analysis used. There are three main types of wAFROC analysis: (a) random reader random cases analysis, which would be generalizable to any reader and any chest x-ray, but would require a huge number of chest x-rays in addition to a large number of participants; (b) random reader *fixed* case analysis—here, the results are generalizable to other radiologists, but are limited to the set of cases that were presented in the experiment; and (c) *fixed* reader *fixed* case analysis, in which case the results are limited to the particular participants and to the particular set of cases used in the experiment.

Ideally, the random reader random case analysis would be used. However, we would have needed a much larger number of cases (up to 150) to be presented to our participants to provide enough statistical power for this analysis. Unfortunately, such a large number of cases could not be displayed to our participants in our setting as it would have demanded too much of the trainees' and radiologists' time. Therefore, we opted for the *random* reader *fixed* case analysis, which allowed us to achieve enough statistical power. Our wAFROC results are considered generalizable to any novice or expert with similar characteristics as our participants, but must be considered specific to the set of cases that were used in our experiment, rather than generalizable to any other chest x-ray.

Future Directions

This work answers research questions that have not been addressed in prior studies and as is the case with scientific research, raises more questions and sheds light on paths for future work. What follows is a discussion of different areas for further research.

Novel Educational Training Systems in Radiology

Research questions to be explored include queries like: could understanding the weaknesses of novice trainees help in guiding the development of remedial curriculums that

focus on these weak spots? Would developing a "weakness-based tutoring system" that concentrates on common false positives help novices improve their performance? Could a virtual problem-based learning environment be created to help trainees work together to decide which cases are normal and which are not, and why? Alternatively, a tutoring platform could be created to provide individuals with necessary scaffolding and close feedback on mistakes, gradually sequencing cases so that they become more difficult with experience. The benefits of such training system is well backed by education theory (A. Collins, 2006; Ericsson et al., 1993; Lajoie, 2009; Lajoie & Azevedo, 2006) and leading examples have already been developed and utilized in other medical and non-medical domains (Lajoie, 2009). Researchers have been calling for a change in teaching methods in the field of radiology and publishing innovative approaches for a long time. Some of these focus particularly on the chest radiograph (Ajlan, Belley, & Kosiuk, 2011; Feigin, 2010; Robinson, 1998). Future research is needed to explore these options to see how best to improve radiology training efforts.

We found certain types of false positives more common in novices. The list of the most common false positive mistakes can be used for targeted training. Perhaps setting up specific review times where experts coach and scaffold novices on how to distinguish abnormalities in these radiological areas can foster a cognitive apprenticeship on what trainees need to know (A. Collins, 2006). Alternatively, setting up a dedicated problem-based mini-curriculum that focuses on distinguishing normal from abnormal radiological appearance of the items on this common false-positives list (Hmelo-Silver & DeSimone, 2013) could also help trainees achieve competence more readily.

Expanding to Non-Radiologists

Quantifiable measures of expertise and errors can be used in educating physicians and trainees in other specialties that rely on the lateral plain film, such as emergency medicine and

pulmonology. This is important for patients; extracting more diagnostic information out of each lateral plain film could help reduce their exposure to radiation by decreasing the need for cross sectional imaging studies such as CT scans. Gaining more out of the frontal and lateral plain films could also decrease the overall cost of health care since plain radiographs are much cheaper compared to cross sectional imaging.

Expanding to Other Radiological Images and Subspecialties

The current study is centered around chest x-rays. These are important and essential as diagnostic tools; however, many other diagnostic modalities and radiologic specialties require similar attention. Future work could extend this research to other modalities, building upon our methodologies and results, and hopefully overcoming the limitations we encountered. This would be important since expertise is domain specific (Chi & Glaser, 1988; Nodine & Mello-Thoms, 2010). An expert in one field, modality or imaging study could be a novice in another, and a high-yield heuristic approach in one imaging modality might fail in another. Every area explored will likely yield new information on the specific false positives or expert mental checklists, for example, in that particular area, which can be translated into rich information for educative purposes, and ultimately lead to optimal patient care. For instance, we discovered important thought pathways experts used that led them to identify a radiological finding through our think aloud component. Future work can examine for other pathways experts follow to discover difficult radiological findings on chest x-rays, or even in other imaging modalities such as CT scans and MRI examinations.

Scan Paths and Gaze Patterns Analyses

The data collected while conducting this research is vast and provides a good database for further post-hoc analysis to expand the results. For example, an in-depth analysis could be done to assess the number of gaze-bounces between the frontal and the lateral views. This could help

understand how important it is to correlate an abnormality seen in one view to the other, before deciding that it is abnormal. Do novices or experts score higher bounces between the views? Which group spends more time correlating with the contralateral view before deciding on diagnosis? Do experts always correlate with the lateral view or are they comfortable calling an abnormality after finding it on a single view? A frontal–lateral correlation analyses could be extended to examine if certain radiologic abnormalities do not require confirmation on the contralateral view; for example, would a fractured vertebral body seen on the lateral require confirmation on the frontal view?

Another possible post-hoc analysis could investigate the number of related abnormalities each group scores, looking for possible patterns. For example, a mediastinal mass is associated with tracheal shift and a smaller lung volume. Would experts score higher on related findings, and would they demonstrate high-domain-knowledge problem solving skills? In other words, do experts look for and find more diagnostic cues in order to narrow down their differential diagnosis list and gain higher confidence about their final diagnosis?

Confidence and Expertise

Another possible post-hoc analysis is to look at participants' confidence ratings of TP and FP lesions, and to plot them against their FOM. Would a Dunning–Kruger effect be demonstrated, so that low performers have a higher confidence about their decisions compared to intermediate performers? Would fellows and staff show a higher confidence compared to others? Or is it part of the nature of reading chest x-rays—that the experts always demonstrate an element of doubt about their findings (Kruger & Dunning, 1999).

Satisfaction of Search

Since many abnormal cases have multiple abnormalities, the dataset provides a good resource to study the satisfaction of search (SOS) phenomenon, where a radiologist misses single

or multiple other abnormalities in a radiologic examination after discovering one abnormality. This phenomenon is known in the literature of medical image perception. First described by Tuddenham (1962), this phenomenon is a hot topic that is still being investigated today (Cook, 2017; Krupinski, Berbaum, Schartz, Caldwell, & Madsen, 2017). Our study can shed some light onto this phenomenon to see if individuals miss findings on the lateral view when they notice an abnormality on the frontal view. Such satisfaction of search question does not seem to have been examined in this type of data.

Think Aloud

Our think aloud results showed that experts and novices *enumerated* a similar number of items when they viewed a normal chest x-ray. One area for future research could be to perform a further post-hoc analysis of the think aloud to understand the similarities and differences between the *types* of items mentioned by each group.

With regard to the pathways verbalized by our participants that led them to identify the mastectomy (Table 1), we were not able to question and investigate the underlying mechanism behind the third pathway (directly calling a mastectomy). This is because of the nature of the think aloud analysis which mandates minimal interference with participants during the think aloud session. Future research can use different qualitative methods tailored to comprehensively investigate the rationale behind direct calling, for example by using a semi-structured interview style.

Conclusion and Summary

This research aimed to identify differences between novices and experts while reading frontal and lateral chest x-rays. The study highlighted some interesting results; some were expected and intuitive while others were novel and informative.

Novices were found to have lower diagnostic performance in interpreting the cases in this study. They were also found to take longer times to decide whether chest x-rays were normal or not. Interestingly, novices and experts did not differ significantly in terms of comfort level reading the lateral chest x-ray, nor in the total number of structures they examine while looking at a chest x-ray. At the same time, novices and experts differed in the types of false positive mistakes they make on these exams.

The think aloud analysis demonstrated important differences between novices' and experts' thought processes and highlighted interesting patterns in each group. Most importantly, more experts tended to search for related findings when they identified an abnormality, which demonstrated their deeper clinical knowledge, pattern recognition, and understanding of associations of findings.

Becoming an expert is a long journey that requires patience, guidance, and deliberate practice. Clear, well-defined landmarks act as safeguards for novice learners so that they do not waste their efforts heading in the wrong direction. Overall, the findings in this study can guide in forming trajectories for novice learners and help them avoid the types of errors we identified. Our hope is that this study will facilitate novices' efforts in reading frontal and lateral chest x-rays and finally achieving expertise.

REFERENCES

Ajlan, A. M., Belley, G., & Kosiuk, J. (2011). "V V O I": A swift hand motion in detecting atelectasis on frontal chest radiographs. *Canadian Association of Radiologists Journal, 62*(2), 146-150. doi:10.1016/j.carj.2010.03.004

Azevedo, R., Faremo, S., & Lajoie, S. P. (2007). Expert-novice differences in mammogram interpretation. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 65-70). Nashville, TN: Cognitive Science Society.

Berbaum, K. S., Franken, E., Caldwell, R. T., & Schartz, K. M. (2010). Satisfaction of search in traditional radiographic imaging. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 107-138). Cambridge, UK: Cambridge University Press.

Berlin, L. (1986). Malpractice and radiologists, update 1986: An 11.5-year perspective. *AJR American Journal of Roentgenology, 147*(6), 1291-1298. doi:10.2214/ajr.147.6.1291

Berlin, L. (1996). Malpractice issues in radiology. Perceptual errors. *AJR American Journal of Roentgenology, 167*(3), 587-590. doi:10.2214/ajr.167.3.8751657

Berlin, L., & Berlin, J. W. (1995). Malpractice and radiologists in Cook County, IL: Trends in 20 years of litigation. *AJR American Journal of Roentgenology, 165*(4), 781-788. doi:10.2214/ajr.165.4.7676967

Chakraborty, D. (2010). Recent developments in FROC methodology. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 216-239). Cambridge, UK: Cambridge University Press.

Chakraborty, D. P. (2011). New developments in observer performance methodology in medical imaging. *Seminars in Nuclear Medicine, 41*(6), 401-418. doi:10.1053/j.semnuclmed.2011.07.001

Chakraborty, D. P., & Berbaum, K. S. (2004). Observer studies involving detection and localization: Modeling, analysis, and validation. *Medical Physics, 31*(8), 2313-2330. doi:10.1118/1.1769352

Chakraborty, D. P., & Zhai, X. (2016). On the meaning of the weighted alternative free-response operating characteristic figure of merit. *Medical Physics, 43*(5), 2548-2557. doi:10.1118/1.4947125

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152. doi:10.1207/s15516709cog0502_2

Chi, M. T. H., & Glaser, R. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv-xxviii). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Collins, A. (2006). Cognitive apprenticeship. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 47-60). Cambridge, UK: Cambridge University Press.

Collins, J. (2001). Radiology training: A program director's perspective. *AJR American Journal of Roentgenology, 177*(5), 1009-1010. doi:10.2214/ajr.177.5.1771009

Cook, T. S. (2017). Now you see it, but would you later? Examining the mechanisms of satisfaction of search in the fatigued radiologist. *Academic Radiology, 24*(9), 1055-1057. doi:10.1016/j.acra.2017.06.004

Crowley, R. S., Naus, G. J., Stewart, J., III, & Friedman, C. P. (2003). Development of visual diagnostic expertise in pathology—an information-processing study. *Journal of the American Medical Informatics Association*, *10*(1), 39-51. doi: 10.1197/jamia.M1123

Delrue, L., Gosselin, R., Ilsen, B., Landeghem, A. V., de Mey, J., & Duyck, P. (2011). Difficulties in the interpretation of chest radiography. In E. E. Coche, B. Ghaye, J. de Mey, & P. Duyck (Eds.), *Comparative interpretation of CT and standard radiography of the chest* (pp. 27-49). Berlin, Germany: Springer.

Ericsson, K. A. (2006a). An introduction to *The Cambridge handbook of expertise and expert performance*: Its development, organization, and content. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 3-20). Cambridge, UK: Cambridge University Press.

Ericsson, K. A. (2006b). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 223-241). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge, UK: Cambridge University Press.

Ericsson, K. A., Krampe, R. T., & Teschromer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363-406. doi:10.1037//0033-295x.100.3.363

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 1-38). Cambridge, UK: Cambridge University Press.

Feigin, D. S. (2010). Lateral chest radiograph a systematic approach. *Academic Radiology, 17*(12), 1560-1566. doi:10.1016/j.acra.2010.07.004

Fitzgerald, R. (2001). Error in radiology. *Clinical Radiology*, *56*(12), 938-946. doi:10.1053/crad.2001.0858

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research, 3*(4), 430-441. doi:10.1177/104973239300300403

Gaber, K. A., McGavin, C. R., & Wells, I. P. (2005). Lateral chest X-ray for physicians. *Journal of the Royal Society of Medicine, 98*(7), 310-312. doi:10.1258/jrsm.98.7.310

Gegenfurtner, A., & Seppanen, M. (2013). Transfer of expertise: An eye tracking and think aloud study using dynamic medical visualizations. *Computers & Education, 63,* 393-403. doi:10.1016/j.compedu.2012.12.021

Hamer, M. M., Morlock, F., Foley, H. T., & Ros, P. R. (1987). Medical malpractice in diagnostic radiology: Claims, compensation, and patient injury. *Radiology, 164*(1), 263-266. doi:10.1148/radiology.164.1.3588916

Hillis, S. (2010). Multireader ROC analysis. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 204-2015). Cambridge, UK: Cambridge University Press.

Hmelo-Silver, C. E., & DeSimone, C. (2013). Problem-based learning: An instructional model for collaborative learning. In C. E. Hmelo-Silver (Ed.), *The international handbook of collaborative learning* (pp. 370-385). New York, NY: Routledge.

Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2010). Learning perceptual aspects of diagnosis in medicine via eye movement modeling examples on patient video cases. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. 1703-1708). Austin, TX: Cognitive Science Society.

Johnson, E. J. (1988). Expertise and decision under uncertainty: Performance and process. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 209-228). Hillsdale, NJ: Erlbaum.

Kelly, B., Rainford, L. A., McEntee, M. F., & Kavanagh, E. C. (2018). Influence of radiology expertise on the perception of nonmedical images. *Journal of Medical Imaging (Bellingham), 5*(3). doi:10.1117/1.JMI.5.3.031402

Kok, E. M., Jarodzka, H., de Bruin, A. B., BinAmir, H. A., Robben, S. G., & van Merrienboer, J. J. (2016). Systematic viewing in radiology: Seeing more, missing less? *Advances in Health Science Education, 21*(1), 189-205. doi:10.1007/s10459-015-9624-y

Kondo, K. L., & Swerdlow, M. (2013). Medical student radiology curriculum: What skills do residency program directors believe are essential for medical students to attain? *Academic Radiology, 20*(3), 263-271. doi:10.1016/j.acra.2012.12.003

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134. doi:10.1037/0022-3514.77.6.1121

Krupinski, E. A., Berbaum, K. S., Schartz, K. M., Caldwell, R. T., & Madsen, M. T. (2017). The impact of fatigue on satisfaction of search in chest radiography. *Academic Radiology, 24*(9), 1058-1063. doi:10.1016/j.acra.2017.03.021

Kundel, H. L., & La Follette, P. S., Jr. (1972). Visual search patterns and experience with radiological images. *Radiology, 103*(3), 523-528. doi:10.1148/103.3.523

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology, 116*(3), 527-532. doi:10.1148/116.3.527

Lajoie, S. P. (2003). Transitions and trajectories for studies of expertise. *Educational Researcher, 32*(8), 21-25. doi:10.3102/0013189x032008021

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from avionics and medicine. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 61-83). Cambridge, UK: Cambridge University Press.

Lajoie, S. P., & Azevedo, R. (2006). Teaching and learning in technology-rich environments. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 803-821). New York, NY: Routledge.

Lee, C. S., Nagy, P. G., Weaver, S. J., & Newman-Toker, D. E. (2013). Cognitive and system factors contributing to diagnostic errors in radiology. *AJR American Journal of Roentgenology, 201*(3), 611-617. doi:10.2214/AJR.12.10375

Llewellyn-Thomas, E., & Lansdown, E. L. (1963). Visual search patterns of radiologists in training. *Radiology, 81*, 288-291. doi:10.1148/81.2.288

Manning, D. (2010). Cognitive factors in reading medical images: A survey of cognitive factors and models of medical image interpretation. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 91-106). Cambridge, UK: Cambridge University Press.

Manning, D., Ethell, S., Donovan, T., & Crawford, T. (2006). How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography*, *12*(2), 134-142. doi:10.1016/j.radi.2005.02.003

Manning, D. J., Ethell, S. C., & Donovan, T. (2004). Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *British Journal of Radiology, 77*(915), 231-235. doi:10.1259/bjr/28883951

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283-298. doi:10.1016/S0001-2998(78)80014-2

Nodine, C., Kundel, H., Mello-Thoms, C., Weinstein, S. P., Orel, S. G., Sullivan, D. C., & Conant, E. F. (1999). How experience and training influence mammography expertise. *Academic Radiology, 6*(10), 575-585. doi:10.1016/S1076-6332(99)80252-9

Nodine, C., & Mello-Thoms, C. (2010). The role of expertise in radiologic image interpretation. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 139-156). Cambridge, UK: Cambridge University Press.

Nodine, C. F., & Krupinski, E. (1998). Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. *Academic Radiology*, *5*(9), 603-612. doi:10.1016/S1076-6332(98)80295-X

Pinto, A., & Brunese, L. (2010). Spectrum of diagnostic errors in radiology. *World Journal of Radiology*, *2*(10), 377-383. doi:10.4329/wjr.v2.i10.377

Pinto, A., Caranci, F., Romano, L., Carrafiello, G., Fonio, P., & Brunese, L. (2012). Learning from errors in radiology: A comprehensive review. *Seminars in Ultrasound, CT and MRI*, *33*(4), 379-382. doi:10.1053/j.sult.2012.01.015

Pollard, B. J., Samei, E., Chawla, A. S., Beam, C., Heyneman, L. E., Koweek, L. M., . . . McAdams, H. P. (2012). The effects of ambient lighting in chest radiology reading rooms. *Journal of Digital Imaging*, *25*(4), 520-526. doi:10.1007/s10278-012-9459-5

Robinson, A. E. (1998). The lateral chest radiograph: Is it doomed to extinction? *Academic Radiology*, *5*(5), 322-323. doi:10.1016/S1076-6332(98)80148-7

Sagel, S. S., Evens, R. G., Forrest, J. V., & Bramson, R. T. (1974). Efficacy of routine screening and lateral chest radiographs in a hospital-based population. *New England Journal of Medicine*, *291*(19), 1001-1004. doi:10.1056/NEJM197411072911904

Spring, D. B., & Tennenhouse, D. J. (1986). Radiology malpractice lawsuits: California jury verdicts. *Radiology*, *159*(3), 811-814. doi:10.1148/radiology.159.3.3704163

Subramaniam, R. M., Sherriff, J., Holmes, K., Chan, M. C., & Shadbolt, B. (2006). Radiology curriculum for medical students: Clinicians' perspectives. *Australasian Radiology, 50*(5), 442-446. doi:10.1111/j.1440-1673.2006.01620.x

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312. doi:10.1016/0959-4752(94)90003-5

Tourassi, G. (2010). Receiver oparating charateristic analysis: Basic concepts and practical applications. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 187-203). Cambridge, UK: Cambridge University Press.

Tuddenham, W. J. (1962). Visual search, image organization, and reader error in roentgen diagnosis. Studies of the psycho-physiology of roentgen image perception. *Radiology*, *78*, 694-704. doi:10.1148/78.5.694

Whang, J. S., Baker, S. R., Patel, R., Luk, L., & Castro, A., III. (2013). The causes of medical malpractice suits against radiologists in the United States. *Radiology*, *266*(2), 548-554. doi:10.1148/radiol.12111119

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive Load Theory: Implications for medical education: AMEE Guide No. 86. *Medical Teacher*, *36*(5), 371-384. doi:10.3109/0142159X.2014.889290

APPENDIX: ELECTRONIC QUESTIONNAIRE

Thank you for participating in this study.

Our goal is to study gaze patterns and thought process of residents, fellows and staff while reading plain films of the chest.

We'll start with some demographic questions about yourself and your expertise in chest radiology.

1. How old are you?

   _____

2. Please select your gender?

   - F
   - M

3. Do you require eyeglasses or any other sight corrective measures?

   - Eyeglasses – Wearing them now.
   - Eyeglasses – Not wearing them now.
   - Contact lenses – Wearing them now.
   - Contact lenses – Not wearing them now.
   - I do not require eye sight correction.

4. What is your residency / fellowship year?

   For staff: Number of years of experience since graduation?

   - R2
   - R3
   - R4
   - R5
   - F1
   - F2
   - _____

5. What is your expected / current sub-specialization?

   (If more than one please indicate all)

   _____

6. How many chest radiology rotations have you done up to this date?

   (For staff: number of years of experience in chest radiology)

   _____

7. How many Emergency radiology AND Plain film reading rotations have you done up to this date?

   (For staff: please keep blank)

   _____

8. How comfortable are you in reading FRONTAL chest x-rays?

   ☐ 1 Very uncomfortable

   ☐ 2 Uncomfortable

   ☐ 3 Neutral

   ☐ 4 Comfortable

   ☐ 5 Very comfortable

9. How comfortable are you in reading LATERAL chest x-rays?

   ☐ 1 Very uncomfortable

   ☐ 2 Uncomfortable

☐ 3 Neutral

☐ 4 Comfortable

☐ 5 Very comfortable

10. How comfortable are you in reading Chest CT scans

    (Overall: High res, PE, trauma, etc…) ?

    ☐ 1 Very uncomfortable

    ☐ 2 Uncomfortable

    ☐ 3 Neutral

    ☐ 4 Comfortable

    ☐ 5 Very comfortable

11. What is your top rated chest radiology resource?

    (If not-specified on the list, please type in the last line)

    - Fundamentals of Diagnostic Radiology (Brant and Helms)
    - Primer of Diagnostic Imaging (The Purple Book)
    - Radiology Review Manual (The Big Green Book by Dahnert)
    - Felsons's Principles of Chest Roentgenology.
    - Core Radiology
    - The Requisites (Chest Imaging)
    - Thoracic Imaging Case Review Series.
    - StatDx.
    - The Teaching Files (by Muller and Silva).
    - _____

End of Questionnaire.