Organised by the CCPL
Complex Cognitive Processing Lab
https://ccpl.hosted.york.ac.uk/

MIPS 2022 Abstracts

Observer perception of breast volume asymmetry on photographs of a physical phantom

Haoqi Wang, BS[1,2], Zhale Nowroozilarki, MS[3], Mary Catherine Bordes, BS[2], Jun Liu, PhD[2], Gregory P. Reece, MD[2], Summer E. Hanson, MD, PhD[4], Fatima A. Merchant[1,3], PhD, Mia K. Markey, PhD[1,5]

[1]Biomedical Engineering, The University of Texas at Austin, [2]Plastic Surgery, The University of Texas MD Anderson Cancer Center, [3]Department of Engineering Technology, University of Houston,[4]Section of Plastic and Reconstructive Surgery, University of Chicago Medicine and Biological Sciences, [5]Imaging Physics, The University of Texas MD Anderson Cancer Center

Rationale

Very few women have perfectly symmetrical breasts. Rather, difference in size between the right and left breasts is normal and most women are not concerned by this asymmetry. However, when women undergo breast surgery as part of their care for breast cancer, they often worry that any breast asymmetry arising from surgical treatment may be easily perceived by others. In this preliminary study, we seek to determine the least volume difference between left and right breasts that observers can detect on photographs of a physical phantom of the torso. Our long-term goal is to help surgeons better counsel patients about how breast cancer and its treatment will change their bodies and whether these changes will be perceivable to other people, when clothed.

Methods

In order to simulate the female torso with different breast sizes, a physical phantom was constructed from a flat-chested mannequin with different size adjustable implants (figure 1A). We utilized 23 different configurations of the physical phantom, where the left and right breast volumes differed from 0 cc to 260 cc. Participants were recruited from Amazon Mechanical Turk and social media advertisements. The observer study was conducted on Qualtrics. Participants were asked to share demographic information. Then, they were shown a random sequence of photographs of the physical phantom taken with a Canon T1i DSLR and asked to judge if the sizes of left and right breasts were the same or different. Psychometric functions were utilized to fit the relationship between the proportion of response signal to absolute volume difference.

Results

Data collection is still ongoing as of the abstract submission, but our preliminary analysis indicates that the perception of breast volume asymmetry is associated with the observer's gender identity. 154 female and 84 male participants were included in the data analyses. As shown in Figure 1B, the difference in breast volume at which 50% of women perceived the breast sizes to be different is lower than the volume difference at which 50% of men perceived the breast sizes to be different. For instance, when the volume difference is 110 cc (Figure 1A, and marked with red circle in Figure 1B), approximately 53% of women thought that the breast sizes were different whereas only 45% of men thought that the breast sizes were different. On the other hand, our preliminary data do not show an association between the perception threshold and the observer age. Other demographic variables of potential interest, such as ethnicity and race, have not been assessed yet because of insufficient data across categories.

Conclusions

We present a human observer study that investigates the least volume difference between left and right breasts that observers can detect on photographs of a physical phantom. We have shown that the threshold is associated with the observer's gender identity. We have not found a strong correlation between age and perception of differences in breast volume. Future work will include targeted recruitment of some demographic groups and more filtering of the data collected via Amazon Mechanical Turk.

A B

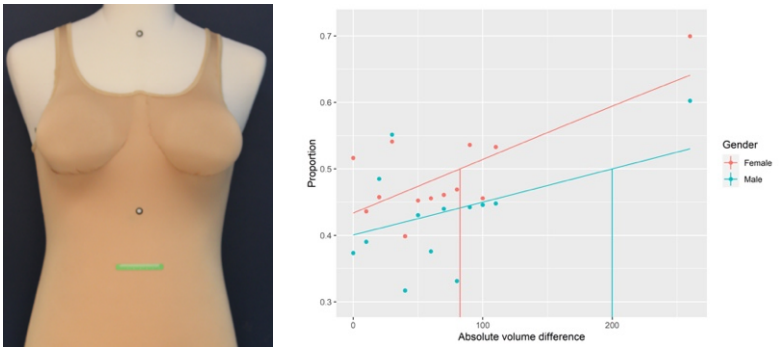


Figure 1. A. Example of a physical phantom of the torso. **B.** Psychometric function fitting. The proportion of the observers that judge the breast sizes to be different increases as the absolute volume difference increases. This analysis also suggests that women perceive differences in breast size more so than do men.

Global Symmetry is Important for the Detection of Abnormality in Mammograms

Cameron Kyle-Davidson, MSc, Emma M. Raat MSc, Lyndon L. Rakusen MSc, Karla K. Evans, PhD
Department of Psychology

Rationale

Expert radiologists can detect abnormalities in rapidly presented mammograms (500 milliseconds) even up to three years prior to onset of cancer (Evans et al., 2019). When radiologists evaluate mammograms, images from the left and right breasts are shown concurrently, given that perception of symmetry between the two breasts is an important indicator of the health of the parenchyma. The importance of symmetry is also evident when radiologists screen for abnormality during rapid presentations since their performance suffers when the bilateral mammograms are from two different women (Evans et al., 2016). In such rapid presentation, there appears to be a global gist signal that abnormality is present and said signal is dependent in part on both mammograms from the patient. Here we investigated whether we could detect with machine learning the global “symmetry” that contributes to the global gist signal and if it affects a pre-trained neural network mammography model.

Methods

We start by investigating whether the global symmetry signal can be detected through a machine learning model. We developed a neural network that accepts the four mammogram views (two per laterality) and predicts whether all four mammograms come from the same woman, or two different women. Mammograms were balanced by size and age to minimise confounding factors. Noting that radiologist performance decreases when the contralateral mammogram is swapped with another woman’s, we tested a pre-trained cancer detection neural network and evaluated its performance with mammograms from same and different women. We used a subset of the Optimam dataset for this task with 5028 Mammograms across 1184 patients.

Results

Our neural network architecture can detect whether a set of mammograms come from the same or different woman with a base accuracy of 61% with fully random swaps. Normalising by age and size increases this ability to 63%. The pre-trained network showed performance differences when mammograms were replaced with those of another woman. There is surprisingly minimal loss in performance when contralateral mammograms are swapped in-category across patients (normal with normal, abnormal with abnormal). Swapping an abnormal contralateral with a normal contralateral mammogram led to a performance decrease but remained better than random chance, as does preserving the abnormal contralateral but swapping the abnormal mammogram itself.

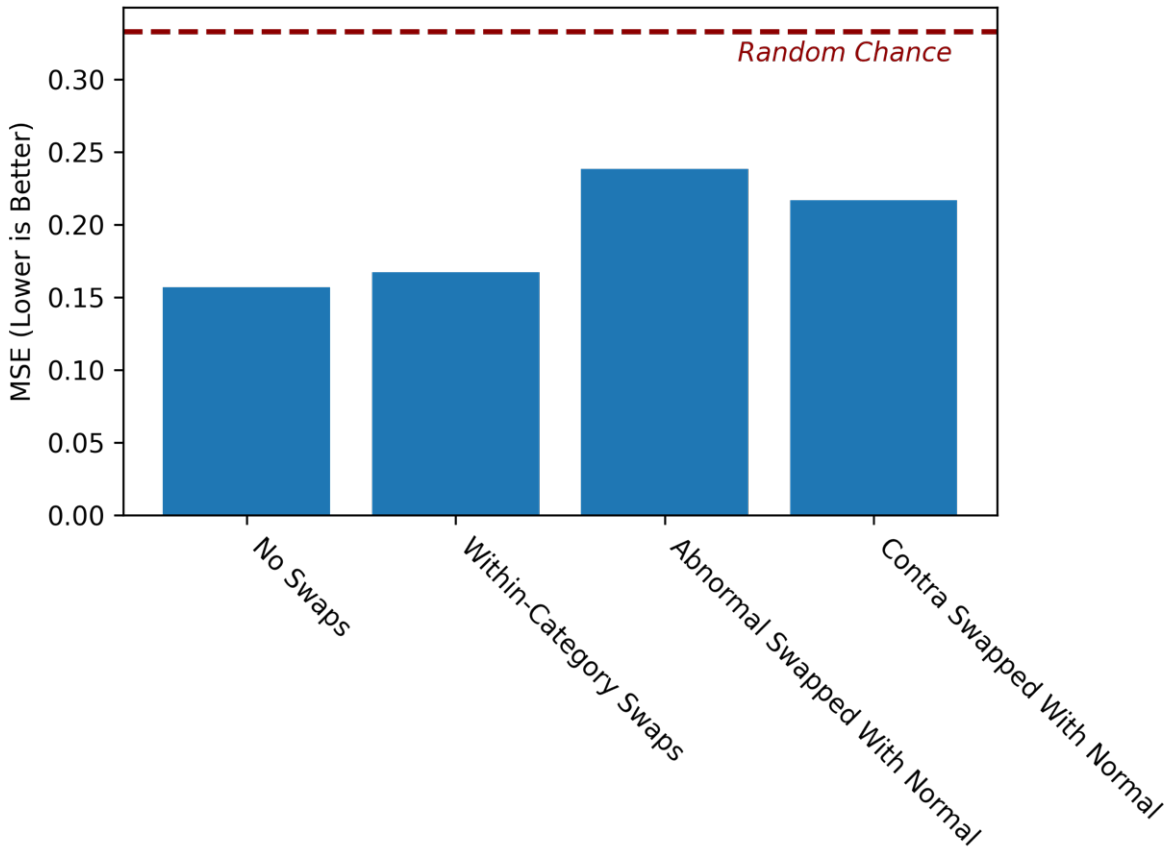


Figure 1: *Cancer prediction error from a neural network in each of the swap cases. Lower is better (more accurate). From left to right: no swapping (All), in-category swapping, abnormal mammogram swapped with normal preserving contralateral, and contralateral swapped with normal contralateral.*

Conclusions

Our neural network can determine whether a set of mammograms comes from the same or a different woman at an above-chance rate. This implies the existence of a global "symmetry" signal common across bilateral mammograms, even when ruling out factors such as breast size and age. This global symmetry signal appears important in the detection of abnormality, given that when a mammogram from one woman is swapped with another, cancer prediction ability falls in machine learning models; even when only the contralateral is replaced with that of another woman. These results closely mirror human performance and might help us understand how global symmetry of the parenchyma is contributing to the global gist of the abnormal, an early cancer indicator in mammograms.

Is it blur, or is it texture?

Andrew William Chen PhD, Murray H. Loew PhD
Department of Biomedical Engineering, The George Washington University

Rationale

If presented with an unblurred picture of a cloud and a blurred picture of a cobblestone road, would one be able to tell which of these images is most in focus? Would a machine? In this work we endeavor to investigate how feature selection may affect the performance of a machine classifier in blur detection. Blur and texture measures are both tools that perform a computation on an image and both are assumed to give useful information about distinct properties of the image. While both types of measures use the same input, the types of conclusions drawn from these measures are very different. We interpret the texture measure (TM) as providing information about only the content of the image before us, i.e., the amount of a particular texture present in an image. The blur measure (BM), however, is interpreted as providing information on some process such as relative motion of the object and imaging apparatus, or defocus of the imaging apparatus, which is explicitly independent of the imaged object.

Methods

We use generalized linear regression models (GLMs) to examine the dependence of BMs on texture information. Three sets of GLMs were constructed from equal-blur-level images of computer-generated mammogram-like clustered lumpy background (CLB) images, as well as from two image sets constructed from the Brodatz texture images. Each image set was designed to contain four distinct textural groups with 100 images per group. **Figure 1** shows the normalized values of TMs and BMs across the four unblurred CLB image sets. The four CLB image sets were called Scaled, Dendrites, Fibrous, and lessBlobs and each of those image types has 100 images. The four image types can be separated by multiple BMs and TMs via one-way ANOVA ($p<0.05$) indicating that the images are of distinct textural groups, as well as of different blur levels though none is blurred. Similar plots can be produced for the two sets of Brodatz images. To refine our models of the dependence of BMs on textural information, we create a set of reduced GLMs containing only those measurement variables (TMs) that are significantly non-zero across all three datasets for each BM. Further, we use five levels of Gaussian blur to artificially blur the CLB images and assess the ability of all of the BMs and TMs used in this study to separate the images in the dataset based on blur level.

Results

We found that many TMs that were used frequently in the reduced GLMs mimicked the structure of the BMs that they modeled. Further, while none of the BMs could separate the CLB images across all levels of blur, a group of TMs could. These TMs occurred infrequently in the reduced GLMs meaning that they rely on information different from that used by the BMs in this study.

Conclusions

Both results should encourage investigators to challenge the classic definitions of these texture, blur, and other measures and to determine which of these measures is appropriate for their specific applications.

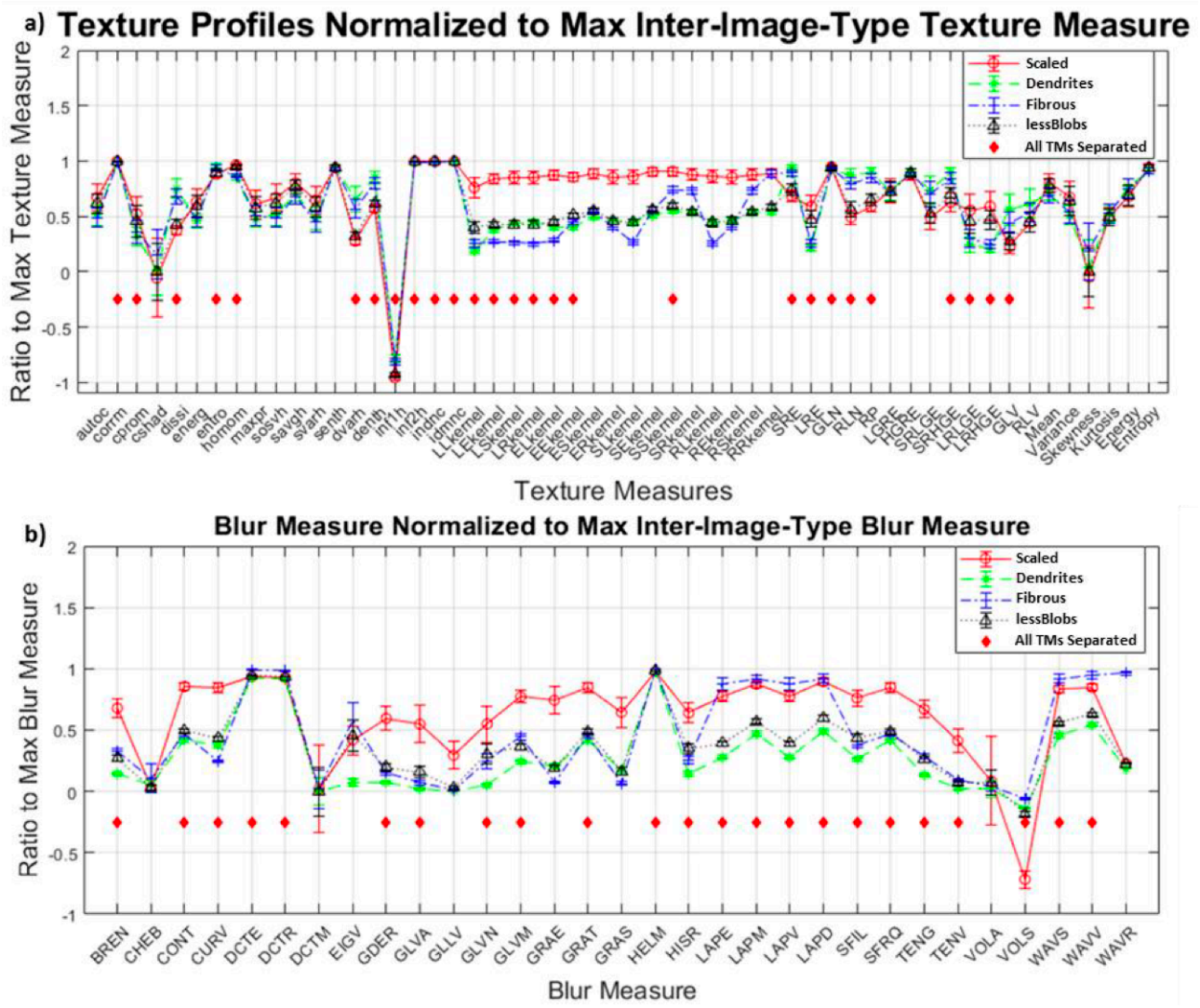


Figure 1. a) Normalized means and standard deviation for all TMs on the unblurred CLB texture images of the four texture groups (Scaled, Dendrites, Fiberous, lessBlobs). **b)** Normalized means and standard deviations for all BMs on the four unblurred CLB texture groups. In both plots red diamonds indicate measures that separate all texture groups with significance ($p<0.05$).

Evaluation of Image Quality Assessment Metrics for MR Images

Yueran Ma[1] (MSc), Jean-Yves TANGUY[2] (MD),
Padraig Corcoran[1] (PhD), Hantao Liu[1] (PhD)
[1]School of Computer Science and Informatics, Cardiff University, United Kingdom
[2]Department of Radiology, Angers University Hospital, France

Introduction

Perceptual quality assessment of medical images has become increasingly necessary. For objective assessment methods of medical image quality, many researchers have conducted research at the level of medical imaging instruments and operations, while little research has been conducted at the algorithmic level. In contrast, the study of objective assessment methods for natural images has been better developed. Therefore, it is imperative to extend the objective assessment methods for natural images to the medical image domain.

Methods

In order to understand the medical image quality perception of radiologists, we conducted a new subjective experiment in a radiology reading room environment at Angers University Hospital Center, France. In this study, image quality was rated by 13 radiologists. In our MRI database, common artefacts were simulated and classified into four categories by whether they were structured and whether the spectral density was flat. In addition, our database contains eight sets of images from six areas of the human anatomy, and also includes two energy levels of artefacts (i.e., physical measure of strength). Therefore, the MRI database consists of a total of 112 medical images (i.e. 8 originals × 2 energy levels × 7 distortion versions).
We evaluated the performance of a total of 32 state-of-the-art traditional (i.e., as opposed to learning-based) objective IQA methods including 29 Full-Reference methods and 3 No-Reference methods based on our MRI database.

Results

The Pearson Linear Correlation Coefficient (PLCC) is used as a measure to assess the performance of these methods. Since the scores produced by objective image quality assessment methods are usually not linear, a nonlinear regression step is essential before the computation of PLCC. The final results are shown in Fig. 1. The top nine FR methods have an accuracy of over 80 percent on MR images, with the best methods approaching 90 percent accuracy, while the NR method does not seem to perform well.

Conclusions

In summary, we conducted a new subjective experiment on medical images using our MRI database and evaluated the state-of-the-art traditional image quality assessment methods on medical images in our MRI database. It is feasible to apply the methodological ideas of natural image quality assessment to medical images and there is certainly room for improvement.
Our results not only give a ranking of the advantages and disadvantages of traditional image quality evaluation methods for medical images, but also shed some light on the differences in the details of feature extraction in medical and natural images. This will help us to propose new objective quality evaluation methods for medical images in the future. We have yet to evaluate the learning-based image quality assessment methods. This will be the focus of our future work.

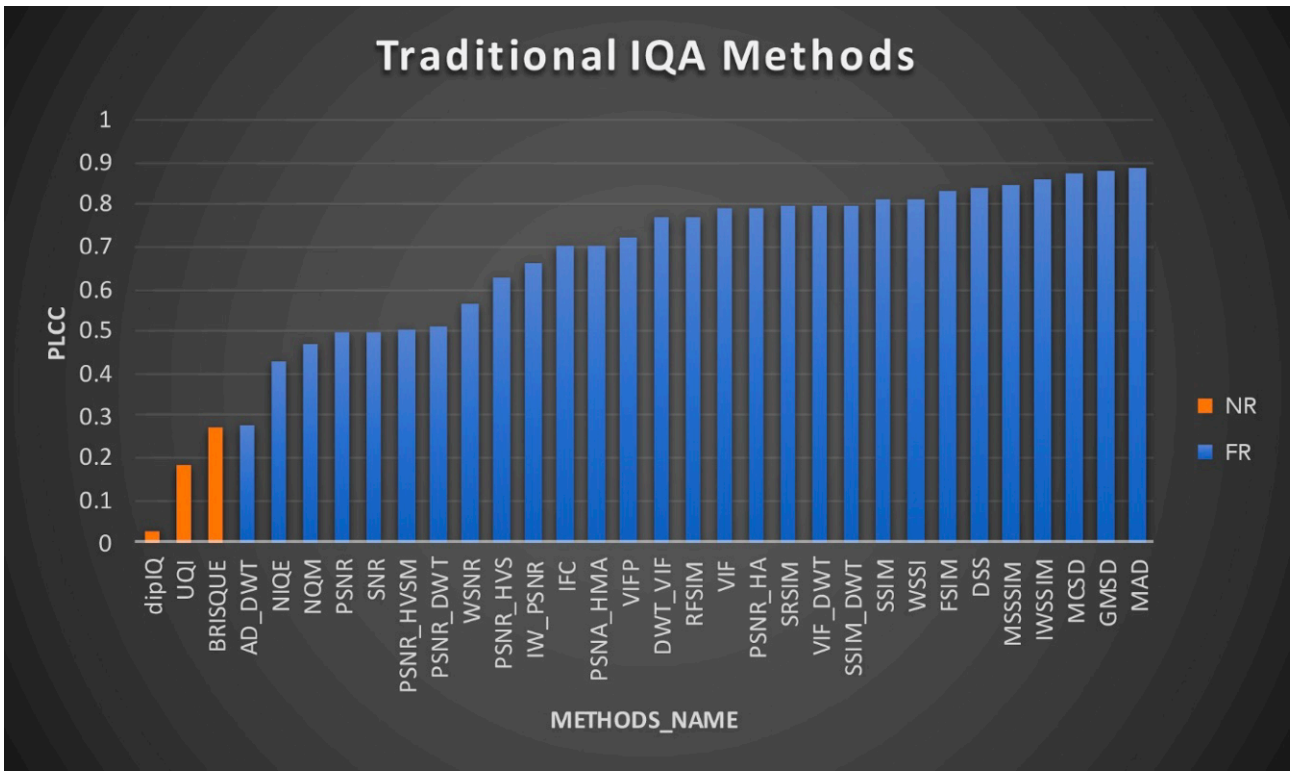


Fig. 1: Evaluation of traditional IQA methods based on our MRI database. Orange bars represent the PLCC of those NR methods, while blue bars represent the PLCC of those FR methods.

Image Quality Assessment of Advanced Reconstruction Algorithm for Point-of-Care MRI Scanner System

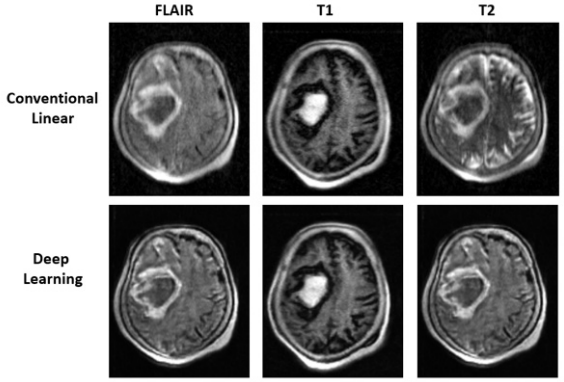
Elizabeth A. Krupinski, PhD[1], Michal Sofka, PhD[2], Jo Schlemper, PhD[2]
[1]Department of Radiology & Imaging Sciences Emory University
[2]Hyperfine operations, Inc.

Rationale

Image quality is crucial for radiology and for radiologists to render efficient and accurate diagnoses. Portable MRI systems have significant potential to rapidly acquire images at the patient's bedside and to improve access in locations currently lacking MRI devices. The scanner under consideration (Swoop®, Hyperfine Operations, Inc., Guilford, CT) has a magnetic field strength of 0.064T (versus 1.5T for most conventional MRIs), thus developing image-processing algorithms to improve image quality is required. This study evaluated images produced using a deep learning-based reconstruction scheme to improve image quality by reducing image blurring and noise to determine if image quality was better than images reconstructed with conventional (linear) methods.

Methods

Six radiologists viewed a randomized series of 90 brain MRI cases (30 acute ischemic stroke, 30 hemorrhage, 30 no lesion) with T1, T2 and FLAIR sequences. Each presentation had both the baseline (linear reconstruction, LR) and deep learning-based advanced reconstruction (AR) images. The task of the observers was to rate the image on the right (AR) as far better, better, same, worse, or far worse than the image on the left (LR) for 3 quality metrics (noise, sharpness, overall) and as strongly agree (no difference), agree (subtle differences but will not affect diagnostic output), neutral (minor differences but unlikely will affect diagnostic output), disagree (modest differences likely will affect diagnostic output), or strongly disagree (large differences will affect diagnostic output) for 3 consistency-based metrics (contrast, geometric fidelity, artifacts). Data were analyzed using Analysis of Variance with metric as the dependent variable and abnormality and sequence as independent variables.



Results

For all metrics there were significant differences as a function of sequence, but not for abnormality. Only artifact revealed a significant difference due to abnormality with acute ischemic stroke having worse artifact scores than hemorrhage and no abnormality cases. There was no consistent pattern in terms of which sequence was worse/better than the others across metrics.

Conclusions

Overall the results indicate that the deep learning-based advanced reconstruction scheme did improve image quality, although degree differs as a function of sequence but not abnormality. Limitations include the fact that only brain images with 2 types of abnormalities were included so results may not generalize to other types of exams. Next steps are to conduct an observer performance study to assess impact of detection accuracy and speed.

Report on NCI's Cognition and Medical Image Perception Think Tank

Todd S. Horowitz, Ph.D.[1], Melissa Treviño, Ph.D.[1,2], George Birdsong, M.D.[3], Ann Carrigan, Ph.D.[4], Peter Choyke, M.D.[5], Trafton Drew, Ph.D.[6], Miguel Eckstein, Ph.D.[7], Anna Fernandez, Ph.D.[8,9], Maryellen Giger, Ph.D.[10], Stephen M. Hewitt, M.D., Ph.D.[11], Yuhong V. Jiang, Ph.D.[12], Bonnie Kudrick, M.S.[13], Susana Martinez-Conde, Ph.D.[14], Stephen Mitroff, Ph.D.[15], Linda Nebeling, Ph.D., M.P.H.[1], Joseph Saltz, M.D., Ph.D.[16], Steven E. Seltzer, M.D.[17,18], Behrouz Shabestari, Ph.D.[19], Lalitha Shankar, M.D., Ph.D.[20], Eliot Siegel, M.D.[21], Mike Tilkin[22], Jennifer S. Trueblood, Ph.D.[23], Alison L. Van Dyke, M.D., Ph.D.[8], Aradhana M. Venkatesan, M.D.[24], David Whitney, Ph.D.[25], Jeremy M. Wolfe, Ph.D.[18,26,27]

1. Behavioral Research Program, National Cancer Institute
2. Clinical Research in Complementary and Integrative Health Branch, National Center for Complementary and Integrative Health
3. Deptartment of Pathology and Laboratory Medicine, Emory University School of Medicine
4. Australian Institute of Health Innovation, Macquarie University
5. Molecular Imaging Program, National Cancer Institute
6. Department of Psychology, University of Utah
7. Dept. of Psychological & Brain Science, University of California, Santa Barbara
8. Surveillance Research Program, National Cancer Institute
9. Booz Allen Hamilton
10. Department of Radiology, University of Chicago
11. Laboratory of Pathology, National Cancer Institute
12. Department of Psychology, University of Minnesota
13. Transportation Security Administration
14. Department of Ophthalmology, SUNY Downstate Health Sciences University
15. Department of Psychology, The George Washington University
16. Department of Biomedical Informatics, Stony Brook University
17. Department of Radiology, Brigham and Women's Hospital
18. Department of Radiology, Harvard Medical School
19. Division of Health Informatics Technologies, National Institute of Biomedical Imaging and Bioengineering
20. Cancer Imaging Program, National Cancer Institute
21. Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine
22. American College of Radiology
23. Department of Psychology, Vanderbilt University
24. Department of Abdominal Imaging, University of Texas MD Anderson Cancer Center
25. Department of Psychology, University of California, Berkeley
26. Department of Surgery, Brigham and Women's Hospital
27. Department of Ophthalmology, Harvard Medical School

Rationale

This is a transformational time for medical imaging. New imaging modalities, the shift to digital technologies, and the rise of machine learning and artificial intelligence are changing the way we peer into the human body. However, none of these technologies can reach their potential if they ignore the humans who interpret the images and algorithms. How can we best move the science of medical image perception forward?

Methods

In September 2019, the National Cancer Institute (NCI) convened a multidisciplinary panel of leading radiologists, pathologists, perceptual and cognitive scientists, and representatives from federal government agencies and professional societies (including the National Institute of Biomedical Imaging and Bioengineering, U.S. Food and Drug Administration, U.S. Department of Homeland Security, and the American College of Radiology) for a "Think Tank" to identify actions to advance medical image perception research. Goals of the Think Tank were to 1) identify critical

research questions, 2) discuss how cognitive and perception research can address these questions, 3) identify barriers to transdisciplinary collaborations and propose solutions, 4) define approaches to elevate the profile of cognition and perception research within the medical image community, and 5) outline future goals and strategies to evaluate progress.

Results

Think Tank participants recommended a multidisciplinary, collaborative approach to medical image perception centered on the clinician's perspective. Critical research questions from a clinical standpoint include: What are the effects of information overload on medical decision-making and decision quality? What strategies could be used to efficiently integrate complex information? How can we document and mitigate the burdens of multitasking? How can we intervene to alleviate the effects of fatigue on interpretive accuracy? How do classification errors change as pathology transitions to digital images?

The Think Tank endorsed a "Reverse Translation" paradigm for addressing these questions. The process begins with clinicians identifying a problem in the clinical setting. Researchers then engage in "use-inspired basic-research", extracting a basic question about the underlying cognitive and perceptual issues that can be studied in the laboratory, sometimes with non-expert observers. After working out the basic science, researchers test the resulting hypotheses with clinician observers in the field.

Major challenges for developing medical image perception science include competition for clinicians' limited time, in their roles as both collaborators and research participants, and access to appropriately annotated medical image datasets. Advancing medical image perception research will require action from universities, medical centers, government funding agencies, and industry to facilitate active multidisciplinary participation from clinicians, perception researchers, and the medical imaging community. Access to datasets can be enhanced by promoting the use of standardized templates for research collaboration and sharing agreements, as well as collaboration among perception science, AI, and medical imaging to appropriately annotate images, ensuring that variables important to perception studies are coded.

Conclusions

Advancing medical image perception research requires active multidisciplinary participation from clinicians, researchers in neuroscience, cognition, and perception, and the medical imaging community. Together these fields can improve patient care by addressing the many diverse challenges facing clinicians today.

Recent methodology developments for analysis of multi-reader multi-case (MRMC) diagnostic radiology studies

Stephen L. Hillis & Brian Smith
Departments of Radiology and Biostatistics

Rationale

For multireader multicase (MRMC) diagnostic studies involving human readers and cases, it is generally preferred to have conclusions that apply to both the reader and case populations, rather than just to one of the populations, as is the case when a t-test or a Wilcoxon test is used. In this talk I discuss recent developments in MRMC analysis methods that accomplish this goal.

Methods

The main method for achieving this goal is the Obuchowski-Rockette (OR) analysis method. Other methods that have been used include the Dorfman-Berbaum-Metz (DBM) method and the U-statistic method proposed by Gallas. Previously it has been shown that the DBM method is equivalent to a specific application of the more general OR method. Very recently, as discussed in the Results, it has been shown that the Gallas method is also equivalent to a specific application of the OR method. For this reason, most of this talk is about the OR method. I limit the discussion to the comparison of reader performance for two imaging modalities, where reader performance is measured by the area under the receiver-operating-characteristic curve (AUC).

Results

Recent developments (in the last five years) for the OR method include the following:

- Establishment of conditions needed for the OR method to produce an unbiased variance
- Establishment of the need for the OR positive constraint on Cov2 – Cov3 to ensure a nonnegative variance estimate
- Establishment that the Gallas approach is a specific application of the OR method without the constraint on Cov2 – Cov3, and thus has the problem that it can result in a negative variance estimate
- Establishment of how to perform region-of-interest (ROI) analyses using the OR method
- Establishment of how to perform analyses involving partially paired data using the OR method
- Establishment of how to define Roe and Metz model parameters that will result in simulated data that emulate any given OR study
- Development of R and Matlab software for implementation of the OR method, including the recent developments listed above.

These developments will be illustrated by real data examples.

Conclusions

These recent developments, coupled with the development of corresponding R and Matlab software, give the researcher more flexibility in designing and analyzing MRMC diagnostic studies.

MRMCaov statistical software for multi-reader multi-case analysis of variance

Brian J Smith, PhD
Department of Biostatistics, University of Iowa, USA
Stephen L Hillis, PhD
Departments of Radiology and Biostatistics, University of Iowa, USA

Rationale

A common study design for comparing the performances of diagnostic imaging tests is to obtain test ratings from multiple readers of multiple cases whose true statuses are known. Typically, there is overlap between the tests, readers, and/or cases for which special analytical methods are needed to perform statistical comparisons. We have developed the MRMCaov statistical software for the analysis of multi-reader multi-case (MRMC) studies of diagnostic tests. In particular, the software enables comparison of reader performance metrics, such as area under the receiver operating characteristic curve (ROC AUC).
R and MATLAB versions of MRMCaov are available and represent first implementations of methods originally proposed by Obuchowski and Rockette (1995) and later unified and improved by Hillis and colleagues (2005, 2007, 2008, 2018). The software is designed to be user friendly and integrated with the respective programming environments. The target audience for MRMCaov is scientists who develop and use diagnostic medical imaging tests, many of whom are R or MATLAB users and attend MIPS. In this talk, we describe features of the software and demonstrate its use with data from an MRMC study with the goal of increasing awareness of MRMC methods and software among the target audience.

Methods

MRMCaov performs MRMC analysis of variance for reader performance comparison of diagnostic imaging tests. The software is open-source with an integrated command-line interface for performing statistical analysis, plotting, and presenting results. It features (1) calculation of reader performance metrics for ROC AUC, likelihood ratio of positive and negative tests, sensitivity, specificity, and expected utility; (2) reader-specific ROC curves; (3) user-definable performance metrics; (4) test-specific estimates of mean performance along with confidence intervals and p-values for statistical comparisons; (5) support for factorial, nested, partially paired, and region-of-interest study designs; (6) inference for random or fixed readers and cases; (7) DeLong, jackknife, and unbiased covariance estimation; and (8) versions for R (https://github.com/brian-j-smith/MRMCaov) and MATLAB (https://github.com/brian-j-smith/MRMCaov.m).

Results

Use of the MRMCaov is illustrated with data from a study comparing the relative performance of cinematic presentation of MRI (CINE MRI) to single spin-echo magnetic resonance imaging (SE MRI) for the detection of thoracic aortic dissection (VanDyke et al. 1993). In the study, 45 patients with aortic dissection and 69 without dissection were imaged with both imaging modalities. Based on the images, five radiologists rated patients on a 5-point scale. Below is code for the MRMCaov R package to estimate ROC curves for each combination of reader and modality and to compare modalities with respect to areas under the curves.

```
## MRMCaov R package analysis of the VanDyke dataset
est <- mrmc(empirical_auc(truth, rating), treatment, reader, case, data = VanDyke)
summary(est)
```

```
## Multi-Reader Multi-Case Analysis of Variance
## Data: VanDyke
## Factor types: Random Readers and Random Cases
## Covariance method: jackknife
##
## Experimental design: factorial
##
## Obuchowski-Rockette variance component and covariance estimates:
##
##                      Estimate Correlation
## reader           0.0015349993          NA
## treatment:reader 0.0002004025          NA
## Error            0.0008022883          NA
## Cov1             0.0003466137   0.4320314
## Cov2             0.0003440748   0.4288668
## Cov3             0.0002390284   0.2979333
##
##
## ANOVA global test of equal treatment empirical_auc:
##
##         MS(T)      MS(T:R)         Cov2         Cov3 Denominator        F df1
## 1 0.004796171 0.0005510306 0.0003440748 0.0002390284 0.001076263 4.456319   1
##        df2    p-value
## 1 15.25967 0.05166569
##
##
## 95% CIs and tests for treatment empirical_auc pairwise differences:
##
##   Comparison    Estimate     StdErr       df      CI.Lower      CI.Upper
## 1      1 - 2 -0.04380032 0.02074862 15.25967 -0.0879594986  0.0003588544
##           t    p-value
## 1 -2.110999 0.05166569
##
##
## 95% treatment empirical_auc CIs (each analysis based only on data for the specified treatment):
##
##    Estimate       MS(R)         Cov2     StdErr       df  CI.Lower  CI.Upper
## 1 0.8970370 0.003082629 0.0004839618 0.03317360 12.74465 0.8252236 0.9688505
## 2 0.9408374 0.001304602 0.0002041879 0.02156637 12.71019 0.8941378 0.9875369
```

Conclusions

MRMCaov brings new tools for MRMC analysis to R and MATLAB and their large communities of users. The software enables comparisons of diagnostic tests within a familiar analysis of variance framework and is notable for the wide range of reader performance metrics and study designs supported. Proper statistical methods and readily available software are crucial for the evaluation and comparison of MRMC studies of diagnostic tests. The MRMCaov software can help ensure the application of such methods.

Investigating Digital Breast Tomosynthesis (DBT) Image Interaction and Associated Eye Behaviours

George Partridge, BSc[a]; Peter Phillips, PhD[b]; Iain Darker, PhD[a]; Yan Chen, PhD[a]
[a]University of Nottingham, School of Medicine, Translational Medical Sciences, Clinical Sciences Building, City Hospital Campus, Hucknall Road, Nottingham, NG5 1PB, United Kingdom; [b]Health and Medical Sciences Group, University of Cumbria, Lancaster, United Kingdom

Rationale

Digital Breast Tomosynthesis (DBT) has been reported to increase reader sensitivity and specificity compared to 2D Digital Mammography (DM) alone. However, due to the increased complexity of DBT images, there's concern that readers may fatigue or hit a cognitive limit sooner which could compromise diagnostic accuracy. Furthermore, due to DBT's relative novelty and due to the challenges of dynamic 3D image eye tracking, there's limited research that characterises reader image interaction and strategies during interpretation. In this study, we will investigate how eye tracked blink behaviours change during a DBT reporting session as proxies for cognitive load and reader fatigue. Additionally, we will begin to analyse image manipulation behaviours along with reader eye gaze to reveal any relationships between performance and reader behaviour.

Methods

A number of UK Breast Screeners (including radiologists and radiographers) were eye tracked as they interpreted and reported on a set of 40 DBT+DM cases with varying difficulty and pathology (47.5% malignant, 12.5% benign and 40% normal) in a random order, as part of the National UK PROSPECTS Trial. Following an initial pilot study, eye tracking data from 35 screeners have been collected from 5 centres in the UK from December 2020 to March 2022, with a further 10 participants from an additional site scheduled in April 2022. Eye tracking data were collected at 60Hz, using non-intrusive eye trackers, along with a workstation screen capture in real-time. All equipment was set up in the readers' natural reading environments at their own sites to replicate normal reading behaviour.

Results

Preliminary analysis found that the blink rate differed significantly between cases when categorised by malignancy (i.e. normal or benign or malignant cases) and outcome (i.e. True Positive, True Negative, False Positive, False Negative), but not with time through reporting session (i.e. as a measure for fatigue). We hypothesised that the changes observed in the blink rate across these different case types could have been indicative of the level of engagement in image search, which in turn could reflect cognitive load. To investigate this further we will relate these data to the image manipulation and eye gaze behaviour that we will analyse in due course. Furthermore, we will progress the analysis of fatigue over the course of the reading session by controlling for other identified potential confounding factors including whether participants took breaks or not and what time of day the reporting session took place. Additionally, we'll investigate the use of the eye blink duration as a potentially more sensitive measure of fatigue in this set up.

Conclusions

This analysis aims to more comprehensively assess significant fluctuations identified in the blink rate over different cases by investigating image manipulation and eye gaze data. Additional analyses of blink behaviour over the time course may also reveal associations between fatigue and diagnostic performance. Findings from this study could highlight the potential use of eye blink behaviours as markers of cognition and fatigue in DBT which could have applications in educational and training contexts, and for advising optimal reading session durations.

Eye tracking ABUS: How radiologists read Automated Breast Ultrasound

Jeremy M Wolfe, PhD [a,b,*], Wanyi Lyu, BA [a], Jeffrey Dong, MD [c], Chia-Chien Wu, PhD [a,b]
[a] Brigham and Women's Hospital, Boston, U.S.A.
[b] Harvard Medical School, Boston, U.S.A.
[c] Beth Israel Deaconess Medical Center, Boston, U.S.A.

Rationale

Automated Breast Ultrasound (ABUS) is a method for producing breast ultrasound imagery in a standardized format. In this, it differs from hand-held ultrasound, where the images are dependent on the sonographer. ABUS creates 3D representations of the breast rather like the stacks of images in CT or MRI. The data are used to construct stacks of images in the coronal and transverse planes. ABUS is especially useful for the assessment of dense breasts. Our goal was to do a 'natural history' study, eye tracking radiologists as they read ABUS cases.

Methods

Twelve readers evaluated three image acquisitions from single-breasts. Radiologists read as many cases as they could in a single 20-minute session. Positive findings were present in 56% of the cases that were examined. We tracked the eyes and monitored the coronal and/or transverse slice that was visible. We used this information to reconstruct 3D "scanpaths".

Results

Readers typically examined all six stacks of images (three coronal, three transverse). Overall accuracy was 0.74 (sensitivity = 0.66, specificity = 0.84, d'=1.4, c=0.3). We classified each miss / false negative error using the Kundel et al (1978) taxonomy of search, recognition, and decision errors. We used cumulative fixation of >1 sec as the threshold for decision errors. Interestingly, of the 20 false negative errors across all readers, 17 are "decision" errors meaning that readers fixated on the target for a second or more but misclassified it as normal or benign. There was just one recognition error and two "search" errors. This is an unusually high proportion of decision errors. Readers had quite stereotyped search strategies that led them to spend about the same proportion of time viewing coronal and transverse images, regardless of the case. In previous work, we identified "scanner" and "driller" eye movement strategies when looking at 3D volumes of image data. Scanners tend to move the eyes widely in the XY plane while moving slowly in Z. Drillers move rapidly in Z while staying fairly fixed in XY before moving the eyes to a new XY location and drilling again in Z. Our readers tended to use a "scanner" strategy for coronal images and a "driller" strategy for transverse images

Conclusions

This high proportion of decision errors suggests that ABUS errors are more likely to be errors of interpretation than of search. Readers found the critical features but did not know what they had found. Further research could determine if readers' stereotyped examination of all stacks of images is useful or if, for instance, readers might triage negative cases on the basis of the coronal images alone, in order to reduce time.

Sequential Reading Effects in Digital Breast Tomosynthesis

Craig K. Abbey1, PhD, Andriy I. Bandos2, PhD, Michael Webster3, PhD, and Margarita L. Zuley2, MD.

1Dept. of Psychological and Brain Sciences, University of California Santa Barbara, CA, USA
2Dept. of Radiology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA
3Dept. of Psychology, University of Nevada Reno, Reno, NV, USA

Rationale

Starting with the work of Taylor-Philips in 2017 [1], a small number of observational studies have reported on the presence of sequential effects in batch reading of mammograms [2, 3]. These studies build on earlier work evaluating batch reading in comparison to reading screening mammograms intermittently with other clinical responsibilities [4, 5]. However, a significant portion of breast-cancer screening has moved to digital breast tomosynthesis (DBT), a quasi-3D imaging modality. DBT imaging appears to result in fewer recalled cases with no penalty in cancer-detection rate, indicating a reduction in false-positive outcomes. DBT also has longer reading times [6], which is an impediment to further adoption of the technology. We do not know of any studies reporting on sequential effects in this relatively new imaging modality. Additionally, the USA has a substantially higher recall rate than the European screening programs that have been reported on to date.

We report initial results of an investigation into sequential reading effects on observational data from the University of Pittsburgh Medical Center (UPMC) where DBT imaging is used routinely for breast cancer screening. We investigate sequential effects on recall and detection rates, as well as reading times.

Methods

Clinical observation data for DBT screening was collected retrospectively from the UPMC radiological information system over 2018-2019, with results from pathology and subsequent imaging up to 2020 and at least 12 months following included screen exams used to establish the cancer status. The initial data consisted of 150,166 records from 17 radiologists. After various exclusion criteria were applied, the resulting data for analysis consisted of 121,652 records, including 1,081 cancers, from 15 radiologists. The data include recall decisions and a final interpretation radiology information system (RIS) timestamp that was used to assess reading time as the difference between timestamps. A difference of more than 10 minutes between sequential timestamps for a given radiologist was used to define a new “batch” of reading data. In this initial presentation of results, we report reader performance measures of false-positive rate, sensitivity, and reading time as a function of the position of a case within a batch. Statistical analysis accounted for between-reader variability (proc glimmix, SAS, v.9.4, SAS Institute, Cary, NC).

Results

Figure 1 shows conditional recall rates and examination time as a function of batch position. The average false-positive rate drops over the first exams in a batch (from 0.16 to 0.11, $p<0.001$), with little evidence of a concomitant drop in sensitivity (from 0.79 to 0.83, $p=0.26$). Reading times are steadily dropping as the batch position increases ($p<0.001$). This drop in false-positives and reading times is consistent with previous reports for mammography [2, 3]. However, the effects are relatively modest by comparison, which may reflect a different imaging modality and a different tolerance for errors in the screening program.

Conclusions

We report initial results from observational data evaluating sequential-reading effects for batch-reading of breast cancer screening exams using DBT. We find evidence for possible early effects in reader performance and a consistent reduction in reading time with batch position.

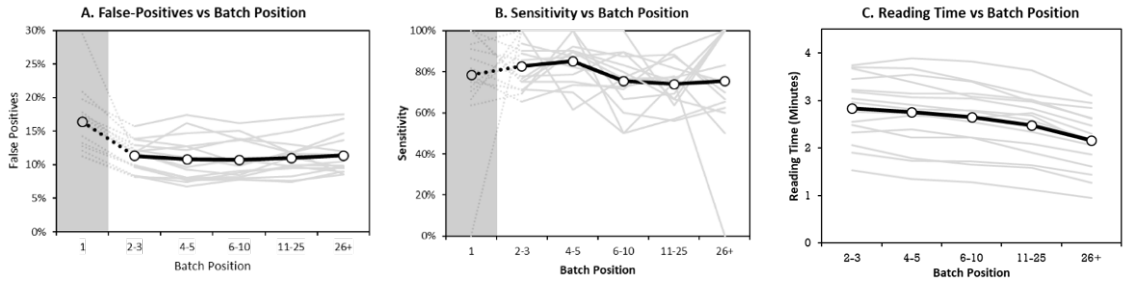


Figure 1. Effects of Batch Position. Reader performance as a function of batch position is plotted for false-positive rate (A), sensitivity (B), and reading time (C). Individual radiologist performance is plotted in gray, with the reader average plotted in black with symbols. There is a considerable spread for individual readers, but appear to be performance differences in the group averages, with decreasing recall rates over the first few batch positions, relatively consistent sensitivity, and decreasing reading times over the range of batch positions. For the reading performance plots (A and B), the first batch position is highlighted (grey background and dotted lines) to indicate that the interpretation of these values may be different than other batch positions because of the way that a batch is defined. Note that there is no meaningful definition of reading time for the first batch position (which is >10 minutes by design), and so it is not plotted (C).

Representing Uncertainty in Visual Diagnosis using Item Response Models

Martin V. Pusic MD PhD[1], Yoon Soo Park PhD[2],
Departments of Pediatrics[1] & Emergency Medicine[1,2], Harvard Medical School

Rationale

Clinicians often find themselves making diagnoses with incomplete information, thus under conditions of uncertainty. It can be difficult to disentangle what is uncertainty due to incomplete mastery and what is structural uncertainty where even a fully trained expert would have difficulty distinguishing between two alternative diagnoses -- where the correct answer would be “could be either diagnosis”. In this work, we explore a statistical approach using item-response models to conjointly quantify diagnostic uncertainty at the clinician and case levels.

Methods

Using an existing dataset, we carried out a prototype analysis of 40 dermatologists rating 100 images of skin lesions that might be diagnosed as malignant melanoma. They used a dichotomous categorization of either “no further treatment” (NFT) or “biopsy/further treatment” (Bx). We modeled the resulting fully crossed data (40 raters x100 pictures of skin lesions) using three approaches: a Rasch Model, a Signal Detection Model (using a reference standard), and a Graded Response Model. We report the decision-thresholds and graphic tracelines for all three approaches and compare them to the individual case locations on the underlying scale.

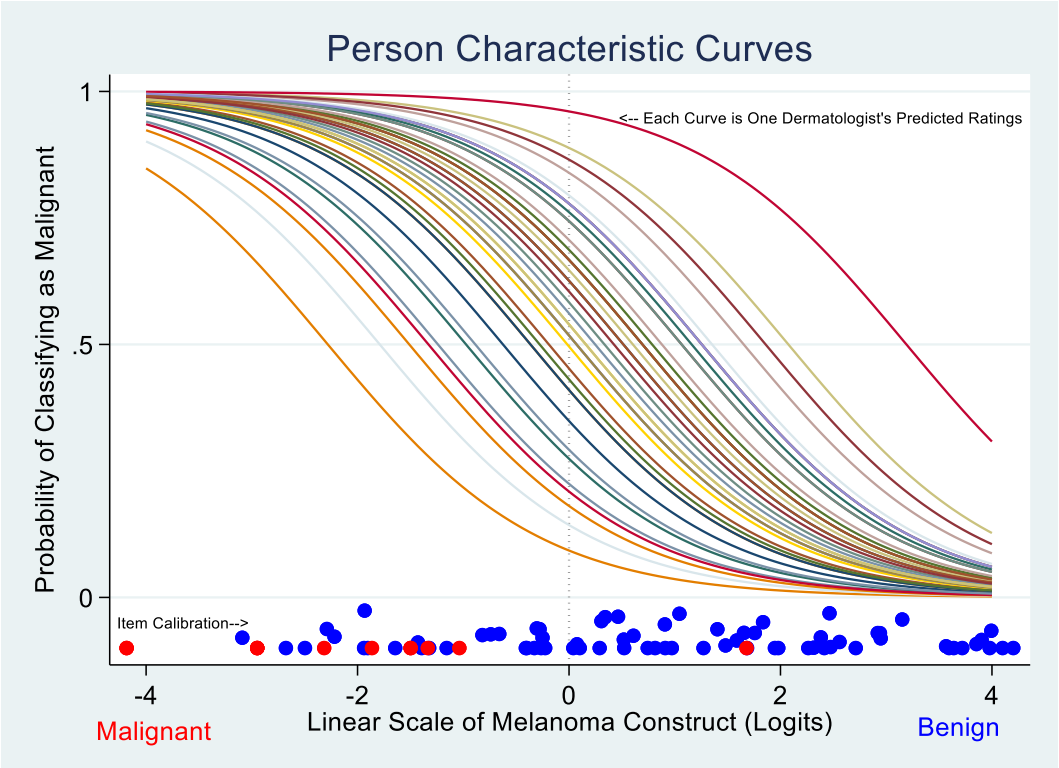
Results

Under the Rasch model, the 100 cases demonstrated a full range of diagnostic uncertainty,
from -4.18 logits (all dermatologists predicted to rate “Bx”) to +4.20 logits (all dermatologists predicted to rate “NFT”). 14 of the cases fell within 0.5 logits of the 0 mid-point where a dermatologist of average bias would be predicted to be equally likely to endorse either category, suggesting there are a considerable proportion of cases with uncertainty for all practitioners. Additionally, several of the ultimately benign cases showed ratings consistent with malignancy. Modelling practitioners, we found that they demonstrated considerable practice variation in where they set their biopsy cut points (See Figure). Signal Detection and Graded Response Model results are found to be complementary to those of the Rasch Model.

Conclusions

Item response modeling, when aligned with an important clinical distinction, such as whether to biopsy or not in a case of potential melanoma, can be used to provide meaningful feedback as to a clinician’s overall tendencies when faced with uncertain cases. The presented work is an advance in that it allows case by case interpretation of an individual’s decision threshold, all taken in the context of demonstrated practice variation.

Figure: Probability tracelines for 40 Dermatologists rating 100 potential melanoma images.



Each traceline is one dermatologist as modeled using a Rasch Model. The case locations are represented by the coloured circles at the x-axis (red=malignant by reference standard; blue=not malignant by reference standard). A case at the zero point would be predicted to be equally likely to be assigned either category (“Biopsy” or “No Further Treatment”) by a clinician of average bias. Clinicians whose traceline mid-point is to the left of zero show a bias towards fewer biopsies; those to the right show a bias towards performing more biopsies.

A Comparison of Conventional Receiver Operating Characteristic (ROC) and Localization-Based ROC (LROC) Analysis

Carl Mauro and Yulei Jiang, PhD
Department of Radiology, The University of Chicago

Rationale

The receiver operating characteristic (ROC) curve depicts possible trade-offs from decision thresholds in imaging search-and-localize and binary classification tasks. However, for search-and-localize tasks, ROC analysis does not naturally account for incorrect lesion localization by readers. “Default-correction” ROC analysis, in which confidence ratings for incorrectly localized lesions are replaced with the lowest (“default”) rating, indicating likelihood of no lesion, is commonly used to account for incorrect localization. “Localization ROC” (LROC) analysis is an alternative method which seeks to achieve more accurate account of incorrect localization. The purpose of this study was to compare conventional ROC analysis with LROC analysis in reader studies.

Methods

Three breast cancer-screening reader studies were analyzed to compare ROC and LROC analyses. Each study involved two different imaging modalities and two or three different localization criteria in multi-reader multi-case (MRMC) analyses: lesion-based (correct localization within 15 mm), laterality-based (correct to laterality), and none (aka case-based). The proper-binormal and non-parametric area under the curve (AUC) were analyzed.

Results

As one would expect, AUC values decreased as localization requirement became stricter. For case-based analysis, proper-binormal and non-parametric AUC estimates are generally comparable. However, under strict localization requirements, these estimates became sometimes highly variable from “default-correction” analysis, but not from LROC analysis. For example, in one study, the difference between proper-binormal and non-parametric AUC estimates increased five-fold from case-based analysis to lesion-based “default-correction” analysis but decreased from LROC analysis (Fig. 1). This was due at least in part to disparate and variable reductions in AUC from case-based to lesion-based “default-correction” analyses (0.19 decrease in non-parametric AUC vs. 0.06 decrease in proper-binormal AUC), whereas the reductions in AUC from LROC analysis remained consistent (0.29 decrease in non-parametric AUC and 0.32 decrease in proper-binormal AUC) (Fig. 1).

Conclusions

Under strict localization requirements, AUC estimates decrease from case-based analysis, and LROC analysis produces similar proper-binormal and non-parametric AUC estimates, allowing proper-binormal estimates to conform more closely to empirical data, thereby characterize reader performance more accurately. Further investigations remain necessary.

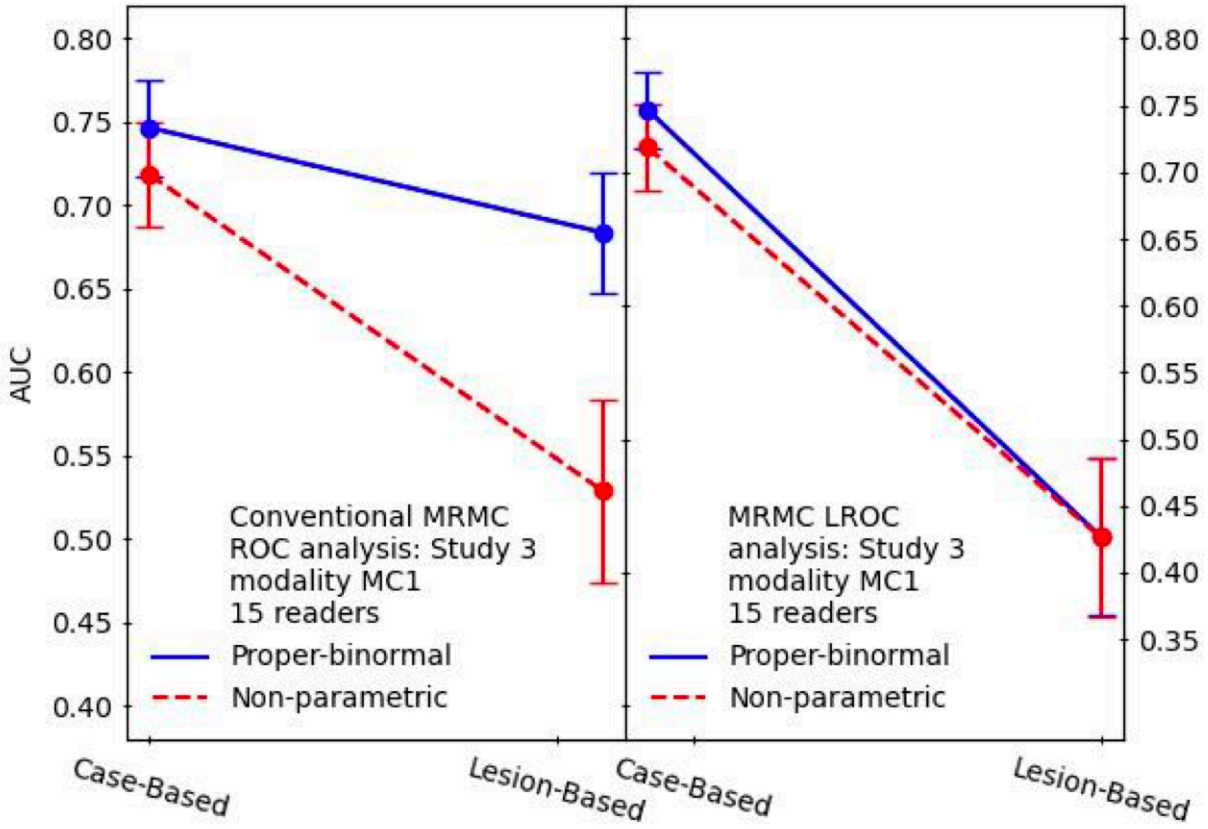


Figure 1. The effect of lesion-localization scoring method on AUC values of individual readers' ROC curves, as estimated by using MRMC analysis. Case-based analysis does not require correct lesion localization and location-based analysis requires correct lesion localization to within 15 mm. Error bars represent the standard error in estimated reader AUC values as calculated using MRMC analysis according to the DBM method.

Evaluating the impact of reader-based and image-based characteristics on diagnostic efficacy in mammography interpretation using a novel analysis method

Dennis Jay Wong, Grad Cert (CE), MDR, BMSc (Hons), Ziba Gandomkar, PhD, MSc, BSc, Warren Reed PhD, Grad Cert (T&L), Bsc (Hons)
Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney

Rationale

To investigate the impact of reader-based and image-based characteristics on reader performance in mammography interpretation to gain greater insight by using novel analysis techniques.

Method

A total of 479 Australian radiologists participated in this study, where six test sets were used. An individual radiologist may have assessed more than one set of the combined 38160 individual case readings. Beforehand these readers completed a survey collecting reader-based characteristics including years in speciality, specialisation, weekly cases interpreted, any fellowship training, and whether the reader work in a breast screening program.
Test sets were from BreastScreen Reader Assessment Strategy (BREAST) and contained 60 cases (20 cancer cases, 40 normal cases), each case included four mammograms: the left and right cranial caudal and medial lateral oblique projections. Image-based characteristics collected included expert difficulty ratings, ground truth, malignancy type, lesion side, site, and size (for positive cases). Case difficulty was determined prospectively by an expert radiologist classifying cases as easy, intermediate, and difficult.

For statistical analysis, a novel linear mixed modelling analysis was used to determine the interactions between reader and image-based characteristics and their significance. Three models were created to evaluate: case-based characteristics; reader-based characteristics; and a combination of reader and case-based characteristics. Normal cases and cancer cases were evaluated separately.

Results

For normal cases, the linear mixed model analysis for image-based characteristics showed expert-determined difficulty was significant ($P<0.001$, 0.004, and $p<0.001$, for easy, intermediate, and difficult, respectively) exhibiting decreasing reader performance as difficulty increased. The reader-based characteristics model found weekly cases interpreted and reader's specialization significant, p-values were <0.001 for 21-60, 61-100,101-150, and >200 weekly cases. For 0-20 cases per week $P=0.003$. Notably, reader performance improved linearly with weekly interpretations but plateaued after 151-200 readings (i.e., annual interpretive volume of approximately 7500 cases). The mixed model's analysis showed a similar trend and significance as the two previous models.

The cancer case's model found none of the image-based characteristics significant. For reader-based characteristics only years in specialty was significant with increasing years improving performance. However, the mixed model found expert-determined difficulty (P-values for easy, intermediate, and difficult were <0.001) significant with increasing difficulty reducing reader performance and number of cases interpreted per week ($P=0.001$, $P=0.001$, $P=0.0002$, $P=0.001$, $P<0.001$, and $P=0.0309$ for 0-20,21-60, 61-100,101-150, 151-200 and >200, respectively) significant.

Conclusion

Expert-determined difficulty and cases interpreted weekly were consistently significant factors impacting reader performance, with increasing difficulty reducing reader performance and increasing cases interpreted improving performance. New analysis methods can provide extra insight into the impact of reader and image-based factors on reader performance however, the

cancer cases' characteristics were inconclusive and require future texture-based radiomic analysis for further investigation.

Consistent performance between experienced and medically naive readers in forced-choice lesion-detection tasks with PET images.

Craig K. Abbey[1], PhD, Sangtae Ahn[2], PhD, Scott D. Wollenweber[3], PhD, Kristen A. Wangerin[3], PhD, Darrin W. Byrd[4], PhD, Fatemeh Behnia[4], MD, Jean Lee[4], MD, PhD, Delphine L. Chen[4], MD, and Paul E. Kinahan[4], PhD.

[1]Dept. of Psychological and Brain Sciences, University of California Santa Barbara, CA, USA
[2]Biology and Applied Physics, GE Research, Niskayuna, NY, USA
[3]GE Healthcare, Waukesha, WI, USA
[4]Dept. of Radiology, University of Washington, Seattle WA, USA

Rationale

Image reconstruction remains an active area of investigation for PET imaging and other tomographic imaging modalities. Reconstruction algorithms balance data fidelity with prior smoothness constraints that control local noise and resolution properties of the reconstructed images. Understanding this tradeoff for optimal task performance is an enduring goal in medical image perception, and a central motivation for the development of model observers.

Machine learning approaches are being considered for the development of network model observers, particularly when the image backgrounds come from patient data with no clearly defined sampling distribution. Thus, there is an ever greater need for labeled training data. Ideally this data would come from radiologists or clinicians that use images in practice. However, it can be a challenge to collect such data because of the limited availability of these expert readers. For simple detection and discrimination tasks that do not require extensive clinical reasoning, non-expert readers might alleviate this constraint in model-observer development. The underlying assumption of this substitution is that these tasks are primarily limited by the perceptual properties of the human visual system rather than the clinical expertise of the readers. However, there has been relatively little testing of this assumption, particularly in PET imaging. In this study, we report on comparisons of clinical and non-clinical readers in a lesion detection study that encompasses reconstruction algorithms, lesion location, and lesion contrast, using list-mode PET data acquired clinically to generate activity backgrounds.

Methods

A series of two-alternative forced-choice experiments with small simulated lesions embedded in clinical backgrounds were conducted with a total of 36 experimental conditions. These included effects of 6 reconstruction algorithms (various parameterizations of OSEM and BSREM), 2 locations (liver and lung), and 3 lesion contrast levels from 162 patient datasets that were used to generate 81 forced choice trials in each condition. The experiments were read by 4 medically naïve subjects, and two radiologists with expertise in PET imaging.
The relatively large number of effects in this data allows for a more thorough evaluation of differences between these two classes of readers. We use two generalized linear mixed effect statistical models to evaluate reader differences statistically, with readers treated either as fixed or random effects.

Results

Figure 1 shows the average performance of expert and naïve readers across reconstruction algorithms. Across all conditions, expert reader performance is slightly lower than the performance of naïve readers (78.5% vs 79.9%), and this contrast is significant in the mixed-effects model with fixed reader effects ($p = 0.0044$). ANOVA modeling for this model finds no significant interaction between readers and signal contrast, image location, or image reconstruction method ($p > 0.05$). The random-reader model finds small variance for reader interactions with lesion contrast and image reconstruction algorithm (< 1% of reader std.), but a more substantial variance for interactions between readers and lesion location.

This study, while small, gives some indication that medically naïve readers may be reliably used in simple forced-choice detection tasks for the purpose of developing model observers in PET imaging.

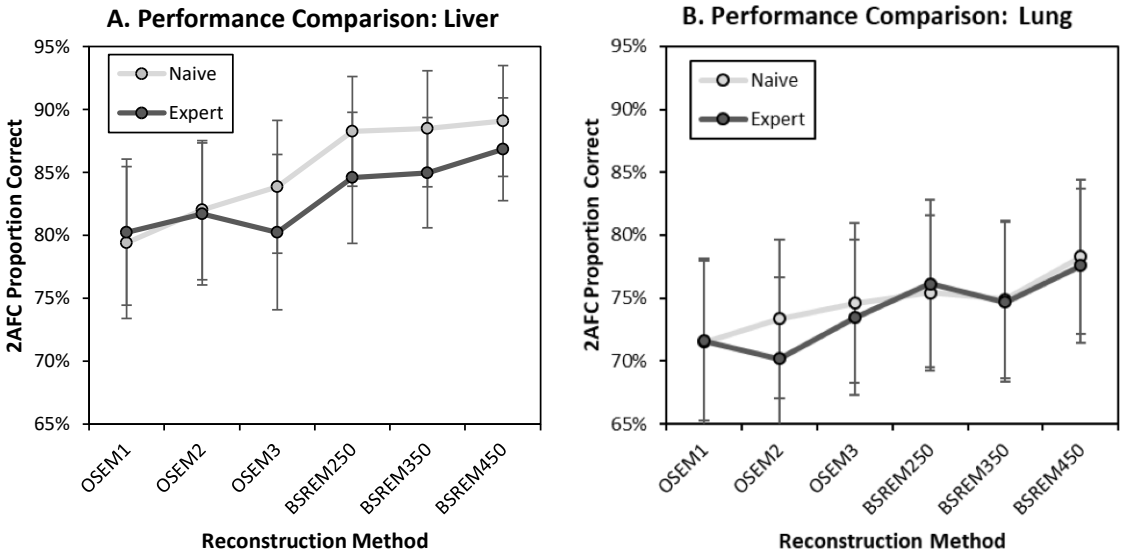


Figure 1. Performance Comparison of expert and medically naïve readers. Comparisons of average 2AFC proportion correct (PC) averaged across the three lesion contrast levels are shown for liver (A) and lung (B) simulated lesions added to clinical PET backgrounds. Liver lesions appear to be more accurately detected than lung lesions, and there appears to be an increasing performance across the 6 reconstruction algorithms (OSEM with 1, 2, 3 iterations, and BSREM with regularization parameters of 250, 350, 450). The medically experienced readers tend to have lower performance than the naïve readers. Error bars represent 95% confidence intervals over the 81 trials and the 3 lesion contrast levels in each experimental condition.

Using computer-simulated nodules to characterize perceptual learning and attentional strategies in lung nodule detection tasks

Frank Tong[1,2], Ph.D., Huiyuan Miao[1], B.S., Hojin Jang[1], Ph.D.,
and Edwin F. Donnelly[3], M.D., Ph.D

[1]Psychology Department, Vanderbilt University
[2]Vanderbilt Vision Research Center, Vanderbilt University
[3]Department of Radiology, Ohio State Wexner Medical Center

Rationale

Both perceptual learning and attentional search strategies are likely important for acquiring and realizing one's radiological expertise, particularly for challenging tasks such as the detection of lung nodules in chest radiographs. In a study performed at the MIP Lab at RSNA 2019, we found that radiologists (N=17) cannot reliably tell apart real nodule cases from simulated nodules in 2D chest radiographs. Nevertheless, individual performance at localizing real nodules was very well predicted by performance with simulated nodules ($r = 0.77$). Here, we evaluated the impact of training non-expert undergraduate participants at nodule detection to better understand how nodule detection performance may depend on perceptually specific learning as well as attentional strategy.

Methods

Observers were trained to detect either visually realistic radiopaque nodules (i.e., light nodules) or polarity-reversed dark nodules in 2D chest X-rays. In recent years, our lab has focused on developing software to generate visually realistic examples of simulated nodules that can vary in shape, size, contrast, and textural properties; these simulated nodules can then be inserted anywhere within the lung cavity of negative chest X-rays to generate simulated lung nodule cases. We have previously reported that undergraduate observers trained with simulated nodule examples demonstrate marked improvements in nodule detection and localization tasks (MIPS, 2019). Here, we assigned 24 observers to each of the two training groups. After undergoing 3 sessions of training with nodules of a given polarity (light or dark), in session 4 observers had to perform separate blocks that required localization of light nodules only, dark nodules only, or search for both types of nodules. Following this experiment, participants performed a nodule localization task involving both real and realistic simulated nodules using a stimulus set that had previously been evaluated on radiologists at RSNA 2017.

Results

Our results revealed much better localization performance for nodules of the trained polarity. Observers trained on light nodules showed much higher localization accuracy when searching for light nodules as compared to dark nodules (86.9% vs. 63.5%, respectively), whereas the opposite pattern was observed for observers trained on dark nodules (73.3% vs. 85.1%). These differences in performance were highly statistically significant. Moreover, when observers were instructed to search for both types of nodules, the performance difference between nodule polarities was further magnified by training condition. Response times indicated somewhat slower performance (by 0.5-1s) when participants had to search for both types of nodules concurrently as compared to search for a single nodule type; however, the increase in response times in the dual target type condition was quite modest when compared to the search times for single polarity nodules. Our results suggest that observers did attempt to perform concurrent search for both nodule types, even if the trained nodule type was attentionally prioritized to some degree. Finally, we found that training with polarity-realistic simulated nodules led to better performance on the RSNA localization test for both real and simulated nodules.

Conclusions

Our findings demonstrate that a perceptually specific template is learned from performing a series of trials requiring lung nodule detection. These learned templates can lead to more effective detection of potential cancer, with performance differences being further amplified by attentional goal or state. Our findings suggest that optimizing the training experiences of radiology residents could potentially lead to better outcomes in diagnostic radiology.

Nodule Hunter: Gamification in Radiology Perceptual Education

Soham Banerjee, MD [1]; Rishabh Agarwal MD [2]; William F. Auffermann, MD/PhD [2*]
[1] Department of Radiology, Baylor College of Medicine, Houston, TX, USA
[2] Department of Radiology and Imaging Sciences, University of Utah Health, Salt Lake City, UT, USA;
[*] Corresponding Author

Rationale

Gamification of education attempts to use the motivational power of games to encourage and enhance learning. Gamers spend a significant amount of time developing and refining game related skills, which teaches them persistence and resourcefulness. Additionally, gamification may potentially offer the ability to teach and assess skills in image perception. While much of traditional education material in radiology focuses on interpretive training, relatively little time is spent on perceptual training. Prior studies have demonstrated promising results using gamification in radiology, but the field remains limited, especially within the field of perception. The goal of this study is to use gamification to teach perceptual skills related to pulmonary nodule identification on chest radiographs (CXRs).

Methods

We created a '3D first person shooter'-inspired videogame called *Nodule Hunter*, which tasks trainees to identify and 'shoot' pulmonary nodules in radiographs. We recruited 8 first year radiology residents and 12 medical students to voluntarily participate in our study. Participants were evenly split into control and experimental groups. Both groups were tasked to identify pulmonary nodules on CXRs using the *RadSimPE* radiology workstation simulator software package, half of the CXRs contained nodules. Then, the experimental group played the game, *Nodule Hunter*, while the control group read an article on intensive care unit imaging. Then both groups identified pulmonary nodules on another set of CXRs. Performance at identifying/marking nodules before and after intervention was compared using receiver operating characteristic (ROC) analysis. We compared area under the curve (AUC) values before and after intervention in both the experimental and control groups. At the end of the study, all trainees were given access to *Nodule Hunter* and the article, and completed surveys describing their thoughts about this experience.

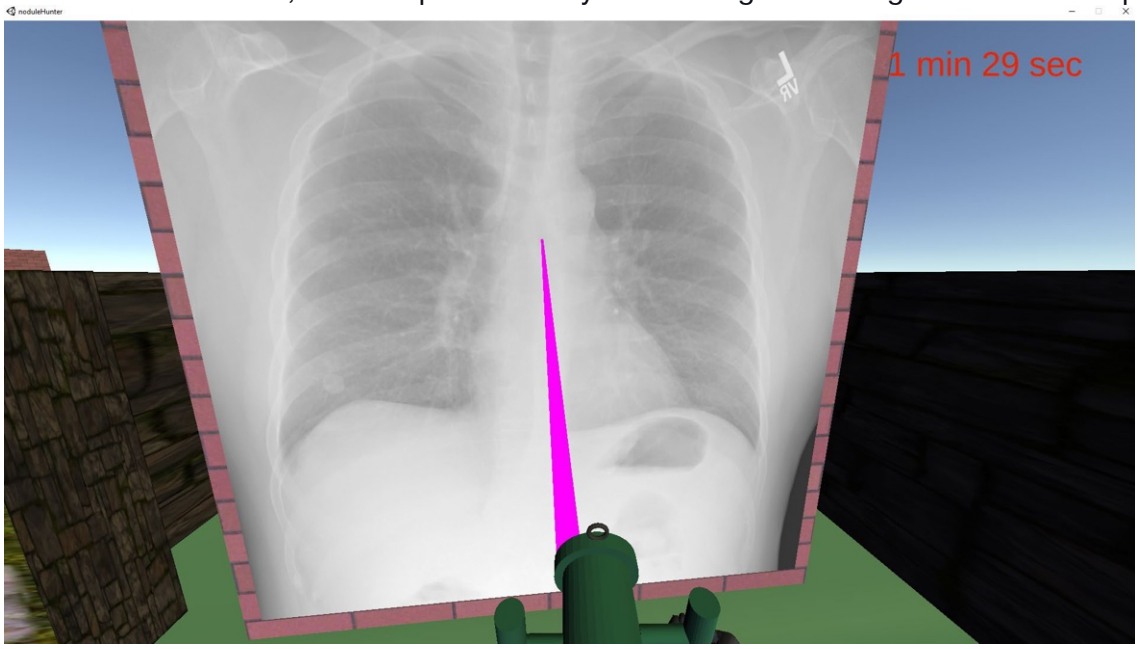


Figure 1: *Nodule Hunter* game. The user needs to identify the right basilar pulmonary nodule in order to continue.

Results

Control and experimental groups both showed a statistically significant improvement in ability to detect nodules based on ROC analysis; Experimental mean ΔAUC = 0.1085, Std = 0.1146, p-value = 0.008; Experimental mean ΔAUC = 0.1780, Std = 0.1038, p-value = 0.0002. Both the control and the experimental group demonstrated statistically significant performance improvement, though the experimental group more so.
Additionally, survey data was substantially positive and statistically significant with all p-values < 0.0004.

Conclusions

We believe that gamification may have an important role in perceptual education, as evidenced by the substantial performance improvement in the experimental group as well as the survey data. Supplementing traditional forms of education, including didactics lectures, reading, and questions with gamification may improve radiology trainee performance.

Cue Utilisation Reduces the Impact of Response Bias in Histopathology

Ann Carrigan, PhD[1,2], Amanda Charlton, MD, FRCPA[3], Mark Wiggins, PhD[4,2] , Andrew Georgiou, PhD[5], Thomas Palmeri, PhD[6], & Kim Curby, PhD[4,2]

[1]Australian Institute of Health Innovation, Macquarie University, Sydney, Australia.
[2]Centre for Elite Performance, Expertise & Training, Macquarie University, Sydney, Australia.
[3]Department of Histopathology, Auckland City Hospital, and Department of Molecular Medicine and Pathology, University of Auckland, New Zealand.
[4]School of Psychological Sciences, Macquarie University, Sydney, Australia.
[5]Centre for Health Systems and Safety Research, Macquarie University, Sydney, Australia.
[6]Department of Psychology, Vanderbilt University, Nashville, United States.

Rationale

Histopathologists make diagnostic decisions that are thought to be based on pattern recognition, likely informed by cue-based associations formed in memory, a process known as cue utilisation. Typically, histopathologists test the working diagnoses derived from cases that have already been classified as 'abnormal' by clinical examination and/or other diagnostic tests. For example, surgical incision for polyps. Consequently, disease prevalence tends to be high. Specimens are also accompanied by a clinical report which likely creates an environment where there is the potential for 'abnormality priming', and a response bias leading to false positives on normal cases. This study investigated whether cue utilisation reduces the influence of a positive response bias in a high disease prevalence context, based on the diagnostic decisions of histopathologists.

Methods

Data were collected from eighty-two histopathologists who completed a series of demographic and experience-related questions and the pathology edition of the Expert Intensive Skills Evaluation 2.0 (EXPERTise 2.0) to establish behavioural indicators of context-related cue utilisation. EXPERTise 2.0 comprises five tasks that are designed to assess participants' ability on: (1) feature identification, (2) feature recognition, (3) feature association, (4) feature discrimination, and (5) feature prioritisation. They also completed a separate, diagnostic task comprising breast histopathology images. To emulate the prevalence rate in a typical non-screening environment (~91%), the participants were presented 123 abnormal cases (malignant and benign) and 12 normal cases for 1500ms and asked to categorise each case. All the tasks were developed with a subject matter expert and co-author (ACh) and took 20-minutes to complete.

Results

Participants were assigned to higher or lower cue utilisation groups based on their performance on EXPERTise 2.0, delineated using a k-means cluster analysis. Forty-seven histopathologists (six trainees) were assigned to the higher group, and thirty-three (fourteen trainees) were assigned to the lower group. As a group, the histopathologists could accurately detect malignant cases, but were less accurate for the normal cases. When the effects of experience were controlled, higher cue utilisation was specifically associated with a greater accuracy classifying normal images ($p < .0167$). We also showed that those with higher cue utilisation recorded a lower positive response bias compared to those with lower cue utilisation ($p = .02$) (See Figure 1).

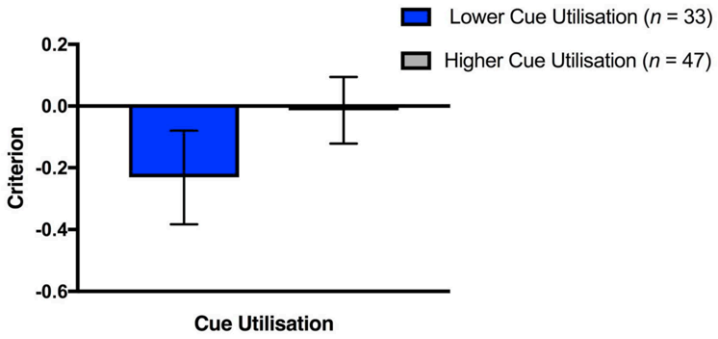


Figure 1. Criterion on the diagnostic task for eighty pathologists, distributed across cue utilisation groups. Errors bars represent 95% confidence intervals.

Conclusions

This findings from this study showed that when controlling for the effects of experience, higher cue utilisation was associated with an increased capability to categorise normal images correctly, resulting in a lower false positive rate. These findings suggest that cue utilisation plays a protective role against response biases in histopathology settings, where the classification of normal tissue is less frequent. The outcomes of the present study have implications for educational initiatives within environments such as histopathology where target or disease prevalence is relatively high, compared with screening environments. Given that current training protocols focus on the identification of disease to support the confirmation of disease, it is crucial that trainees and experienced histopathologists develop and maintain their ability to recognise the range and variability or normal pathology to avoid false positive errors and subsequent harm to patients.

Using Decision-Aligned Response Models to Represent Discrete Categories of a Histological Continuum

Martin V. Pusic MD PhD[1], Amy Rapkiewicz MD[2], Jonathan Melamed MD[2]
[1]Departments of Pediatrics & Emergency Medicine, Harvard Medical School
[2]Department of Pathology, NYU Grossman School of Medicine

Rationale

Pathologists often need to make discrete categorizations based on a continuous underlying latent construct. The example we will consider is the staging of Prostate Cancer Histology into the International Society of Urological Pathologists' five (ISUP) categories, each of which have different subsequent management. Existing statistical measures based on contingency tables -- agreement statistics (e.g., kappa) and Signal Detection Models (AUC) -- do not use all available information. In this work, we use item-response models to conjointly estimate an individual clinician's decision thresholds and their predicted responses for individual cases, with a view to more generalizably reporting practitioner variation in diagnosis.

Methods

We used an existing dataset where an international panel of 24 urological pathologists graded 50 prostate cancer images (Egevad 2018), supplemented with new ratings by 13 local pathology residents. Each participant assigned one of the five ISUP categories to each image; successive categories suggest higher malignant potential. The dataset was modelled using a graded-response model, providing estimates for both cases (location on continuum, standard error) and individuals (location of decision thresholds with standard errors; tracelines). A consensus (16/24 experts) interpretation was available for all cases.

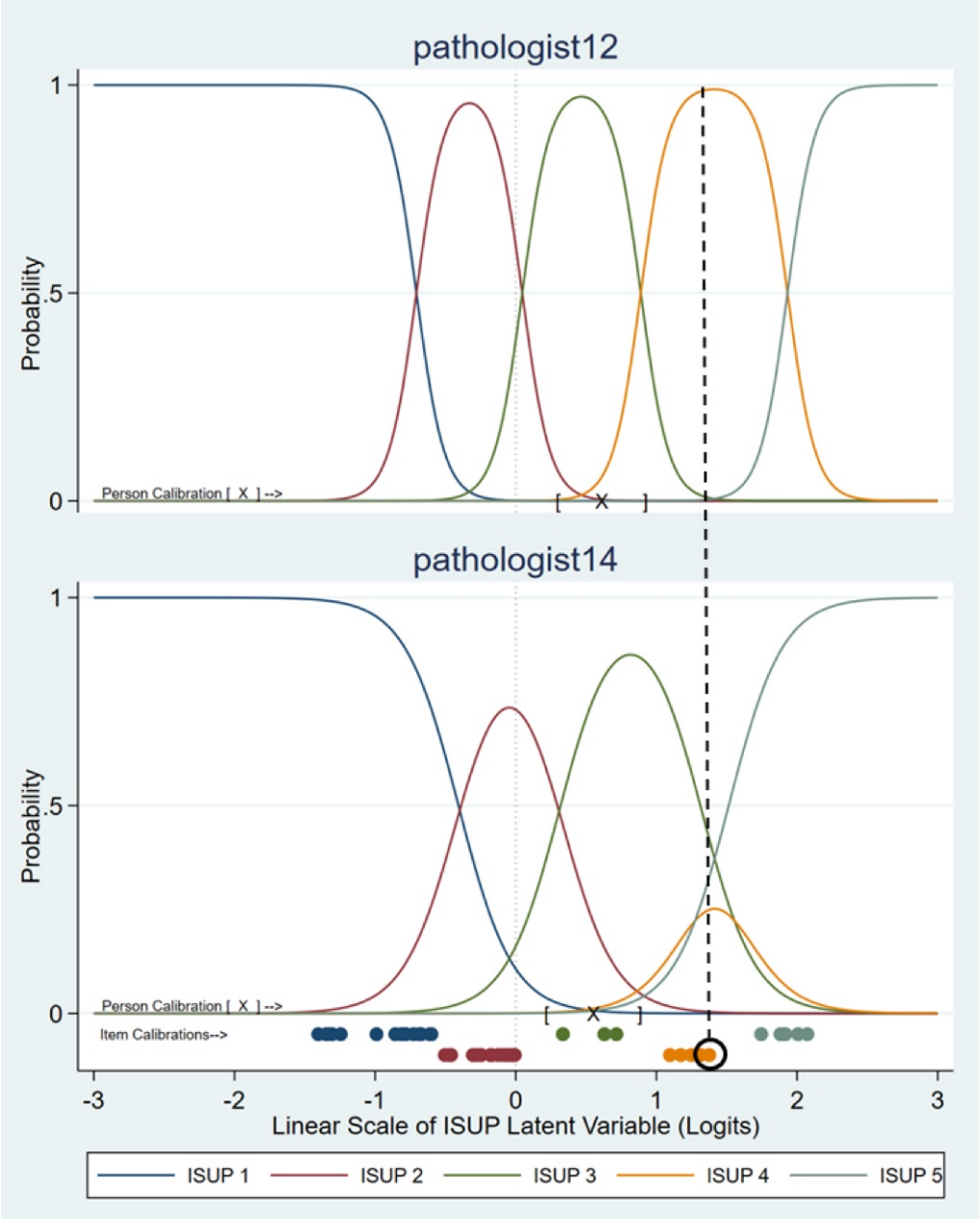
Results

Under the response model, the 50 cases demonstrated the full range of diagnostic severity, with the case locations corresponding well to the consensus standard determinations (Figure, lowest pane, scatter plot). Participants' decision thresholds generally aligned with the items. There was a clear expert-novice difference as well as significant practice variation between expert pathologists, such that even pathologists whose agreement with the reference standard was the same could be shown to have quantitatively different decision-thresholds.(Figure)

Conclusions

When considering a diagnostic continuum, it is possible to model both the severity of the case and the clinician's decision thresholds on the same informative latent scale. This method has the potential to provide feedback to clinicians as to their decision tendencies, customized to specific regions along the diagnostic continuum.

Figure: Probability tracelines for two pathologists rating 50 prostate cancer images



Each traceline represents the diagnostic category most likely to be chosen by the pathologist for a case at that location on the severity scale, as modeled using an Item Response Model. The case locations are represented by the coloured circles at the lower x-axis (colour denotes ISUP consensus diagnosis). Distance of case along the x-axis indicates severity of diagnosis, with the consensus diagnosis and distance aligning very well. We show the five category tracelines of two clinicians with the same level of agreement with the consensus diagnoses (κ=0.72) but whose tracelines show a considerably different pattern. Pathologist 12 shows sharp thresholds between each diagnostic category; Pathologist 14 shows some difficulty with ISUP Category 4. A case with severity at the dashed line would be predicted to be diagnosed very differently by the two pathologists.

Using a Limited Field of View to Improve Nodule Identification on Radiographs

Rishabh Agarwal, MD [1]; Sam Zenger, MD [2]; William F. Auffermann, MD/PhD [1*]
[1] Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, UT, USA
[2] Billings Clinic Internal Medicine, Billings, MT, USA; [*] Corresponding Author

Rationale

Perceptual error is a significant cause of medical errors in radiology. Given the amount of information in a medical image, it may be difficult for an image interpreter to appropriately focus on their search pattern. For example, when searching for pulmonary nodules, parts of the image outside the lung contain potentially distracting information not relevant to the identification of nodules. Also, when viewing an image for pulmonary nodule identification, the regions outside the lungs appear very bright, which may lead to eye strain during prolonged image viewing and result in perceptual errors. The goal of this study is to examine the effects of presenting images with a limited field of view (LFoV) on identification of pulmonary nodules on chest radiographs (CXRs).

Methods

To date, six medical students participated in this IRB approved study and were split evenly into control and experimental groups. Both groups were introduced to our RadSimPE radiology simulator software, the basic aspects of pulmonary nodules, and a lung search pattern. Both groups were shown a first set of 20 CXRs (CS1), of which half contained nodules. Participants were asked to rate the probability of a nodule being present using a receiver operating characteristic (ROC) scale, mark the location of nodules, and indicate their confidence in localization. Then the experimental group was given training on localizing nodules on LFoV CXRs, while the control group received an attentional control journal article. For their second set of cases, the control and experimental groups were shown similar images, the only difference was that the experimental group's LFoV CXRs were masked to exclude the chest wall and abdomen. Participants were then asked to repeat the same image evaluation tasks as for CS1. At the end of the study, participants were given surveys containing 8 items to assess their perception of the training.

Results

There was a significant improvement in performance in nodule identification for the experimental group (mean ΔAUC= 0.1983, StDev= 0.0653, P-value= 0.017), but not the control group (mean ΔAUC= 0.1250, StDev= 0.1085, P-value= 0.0921). There was no significant change in nodule localization accuracy or localization confidence for either group. Survey results were significant for 4 of 8 survey questions. Participants subjectively stated they found the limited field of view images easier to view (less eye strain) relative to standard images.

Conclusions

Using limited field of view images may be a useful technique when performing specific high yield perceptual search tasks.

How many cues does it take to find every cancer?

Dr Damien Litchfield[1], PhD & Dr Tim Donovan[2], PhD

[1]*Department of Psychology, Edge Hill University*
[2]*Institute of Health, University of Cumbria*

Rationale

Understanding how observers interpret complex medical images and detect pathology typically compares between experts and novices to establish what processes are optimised for high performance, e.g., more efficient eye movements (Donovan & Litchfield, 2013) or exploiting the first glimpse of the scene (Litchfield & Donovan, 2016). Yet despite extensive training, experts still miss cancers (~4-20%, Goddard, 2001) and rarely achieve 100% cancer detection in experiments. In this study we took a novel experimental approach using eye-tracking and asked novice observers to detect lung nodules from chest x-rays whilst making use of cues to aid their detection of these targets. Without cues novice accuracy is typically ~50% whereas experts achieve 80%-90% (Donovan & Litchfield, 2013). Our goal was to establish what it would take to achieve 100% detection and whether this was possible just by knowing the precise visual depiction of the target to be found on each image, or if additional cues were required.

Method

60 novice observers were presented 36 chest x-ray images and asked identify single lung nodules within each chest x-ray without time constraints. Throughout each trial a cue to the right of each x-ray showed the precise target that observers needed to find on that image. These target cues consisted of the specific lung nodule target image taken from the medical image and cropped with either a 1-pixel border showing just the target only, or 100-pixel cropped border which also showed some of the surrounding image details further away from the target, and therefore also providing some location cue information. A between participants design randomly allocated participants to either the 1-pixel or the 100-pixel condition. Eye movements were recorded and areas of interest were created around the target in each medical image and also around the target cue to establish search behaviour on the target vs on the cue.

Results

Accuracy in the 1-pixel cue condition was 65% (with 1 novice achieving 100%) whereas participants with 100-pixel cues (including surrounding spatial information), had significantly higher accuracy (86%: with 4 novices achieving 100%). The 100-pixel cue led to significantly faster decisions (5.42sec vs 6.95sec), fewer fixations (11.4 vs 13.4) and less time needed to initially find the target (3.02sec vs .3.55sec) compared to the 1-pixel cue. Verification time (time since first fixating target and making a decision) was also significantly faster with 100-pixel cue than 1 pixel cue (2.4sec vs 3.39sec). Significantly more trial time was spent fixating the cues in the 100-pixel cue (22% vs 18%), but there was no significant difference in the number of cue revisits since fixating the target (1-pix = 1.07, 100-pix = 2.10).

Conclusions

We discuss how observers make use of these cues, how such cues change observer search and why lung nodules are still hard to find even when shown precisely what the target is.

The effect of spatial frequency on gist perception in medical imaging

Emma M. Raat, MSc, Karla K. Evans, PhD
Psychology Department, University of York, England

Rationale

Rapid extraction of global structural regularities provides us with basic information of our visual world, the so-called gist. Scene research is divided as to whether it is low or high frequencies that are essential for driving gist signals from natural scenes. However, recent medical imaging perception work has suggested that high rather than low spatial frequencies are vital for radiologists to extract the gist of abnormality from mammograms (Evans et al., 2016). Here we investigate how maintaining a series of higher spatial frequencies in an image might aid perception of the gist of abnormality across different types of mammograms.

Methods

This study aimed to investigate the effects of high-pass filtering of mammograms on detecting the gist of medical abnormality. Mammograms were viewed for 500 ms, masked, and then rated on a 0-100 rating scale for gist impression of abnormality in three counterbalanced blocks of 180 trials each. The set of mammograms consisted of 90 no cancer, 30 obvious cancer, 30 subtle cancer, and 30 three years prior to sign of abnormality acquired images. Performance on the unaltered set of mammograms (F0) was compared to four levels of contrast-normalized, 2nd order Butterworth high-pass filtered versions of the same cases. A total of thirty experienced radiologists (>1000 scans read last year) took part across 3 variations of the experiment (F0.5 & 1 cycles per degree (cpd); 0.5 & 1.5 cpd; 1 & 2 cpd; 1.5 & 2 cpd).

Results

As expected, there was a strong main effect of image type ($\chi^2(2)=147.51$, $p=<.001$; $\log_{10}(BF_{10})= 21.99$), where performance was highest for obvious (AUC=0.68), then subtle (0.60), visible cancers with reduced performance for priors (0.53) with no visibly actionable lesion across frequencies ($p<.001$ for each comparison). There was also a main effect of frequency ($\chi^2(8)=61.93$, $p=<.001$; $\log_{10}(BF_{10})= 4.31$), where we found no significant difference between F0 and 0.5 cpd or 1.5 cpd filtered mammograms, but a significantly better performance for unfiltered than 1 cpd ($p<.001$) and 2 cpd ($p=<.001$) filtered mammograms across image types. Interestingly, there was also strong evidence for an interaction effect between image type and frequency ($\chi^2(4)=36.836$, $p<.001$; $\log_{10}(BF)= 7.65$). Post-hoc comparisons showed that AUC for priors was significantly higher at 0.5 (0.61) and 1.5 cpd (0.61) than for unfiltered mammograms (0.47) ($p<.001$ for each).

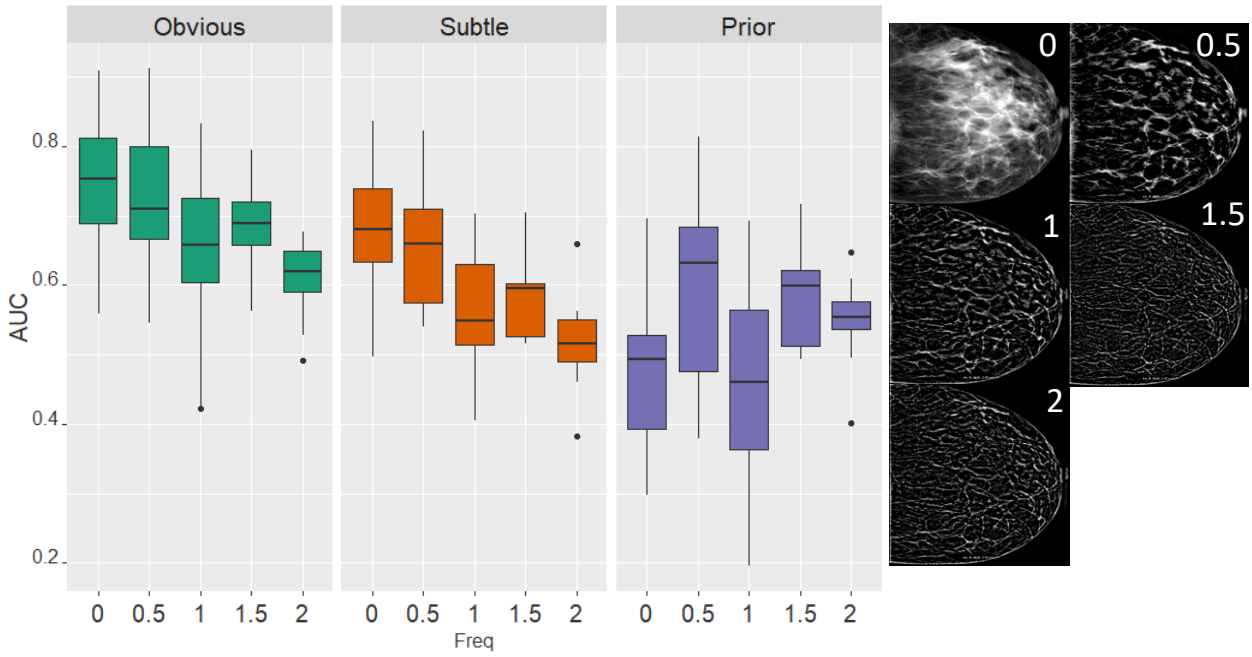


Figure 1: *On the left, boxplots of AUC distribution per conspicuity (obvious, subtle, and prior) and high-pass filter spatial frequency level. On the right, an example mammogram across each of the high-pass filter levels*

Conclusions

In conclusion, the preserved overall performance with 0.5 and 1.5 cpd high-pass filtered mammograms combined with the increased gist perception for mammograms acquired 3 years prior to onset of any visible signs of cancer for the same suggests certain high-pass filters might intensify subtle global textural patterns. These global patterns could play an important role in early cancer detection. Further research using band-stop filters might provide more insight into the importance of specific frequencies.

Lack of global mammographic signature associated with malignancy might cause a false-negative diagnosis

Ziba Gandomkar*, PhD, MSc, BSc, Somphone Siviengphanom BSc, Mo'ayyad Suleiman PhD, MSc, BSc, Dennis Wong, Grad Cert (CE), MDR, BMSc, Warren Reed PhD, Grad Cert (T&L), BSc, Ernest U. Ekpo, PhD, BSc, Sarah J. Lewis, PhD, MEd, BAppSci, Patrick C. Brennan PhD, BSc.

Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney

Rationale

Earlier studies on computer-aided breast cancer detection tools showed global image features extracted from mammograms, or global radiomic signature, can indicate that malignancy appearances are present within an image [1,2]. This study focuses on a set of screen-detected breast malignancies, which were also visible on the prior screening examinations (i.e., missed cancers based on the priors). We explored if the global radiomic signature could differentiate between screening rounds: when the cancer was detected, from the round immediately before. We also explored if signatures from the missed cases were similar to normal images.

Method

Using a retrospective radiological review of screen-detected images, a set of 129 cases, where the cancer was missed on the prior examination was identified. These images were retrieved from an archive of BreastScreen Australia, where standard practice is double reading with the arbitration. Therefore, these malignancies were missed by two original radiologists in the screening program but a panel of three experienced radiologists, who retrospectively interpreted the prior examinations (knowing that a later screening round had revealed a cancer), indicated that the cancer was visible, and its signs were actionable. We asked the panel to exclude any case, where the sign is visible but non-actionable. Both current (i.e., the images on which cancer was detected during the screening program or “ID” category), and prior images (i.e., the images from the previous round of screening or “Missed” category) were collected using a single vendor technology as the appearance of mammograms, and hence, the global image feature could affect the radiomic signature of the images.

A global radiomic signature was extracted from each image. This signature included 12 histogram-based and 22 Haralick texture features. Bilateral craniocaudal (CC) and mediolateral oblique (MLO) views were available for each examination. Two separate random forest classifiers for each view were built. To produce the case-level malignancy probability of a screening examination, the maximum value of malignancy probabilities, assigned to four images (CC and MLO, right and left) corresponding to the exam was considered. To train and validate the model, leave-one case out cross-validation was used. A set of normal cases, matched based on mammography machines, was also retrieved.

Results

As shown in Figure 1, the classifier resulted in an AUC of 0.66 (95% Confidence Interval (CI)=0.60-0.73) for differentiating “Missed” images from “ID” ones. The AUC for differentiating “Normals” from “Missed” and “Normals” from “ID” images were 0.53 (95%CI= 0.48-0.58) and 0.65 (95%CI=0.60-0.69), respectively.

Conclusion

The classifiers’ performances imply that global image signature of “Missed” examination is akin to the signature for “Normals” while varied from the signature of identified cancers. Therefore, considering the importance of global processing in the mammogram interpretation process, eliminating some of these “Missed” cancers would be challenging as the global impressions of malignancy that help with a diagnosis, are at best weak.

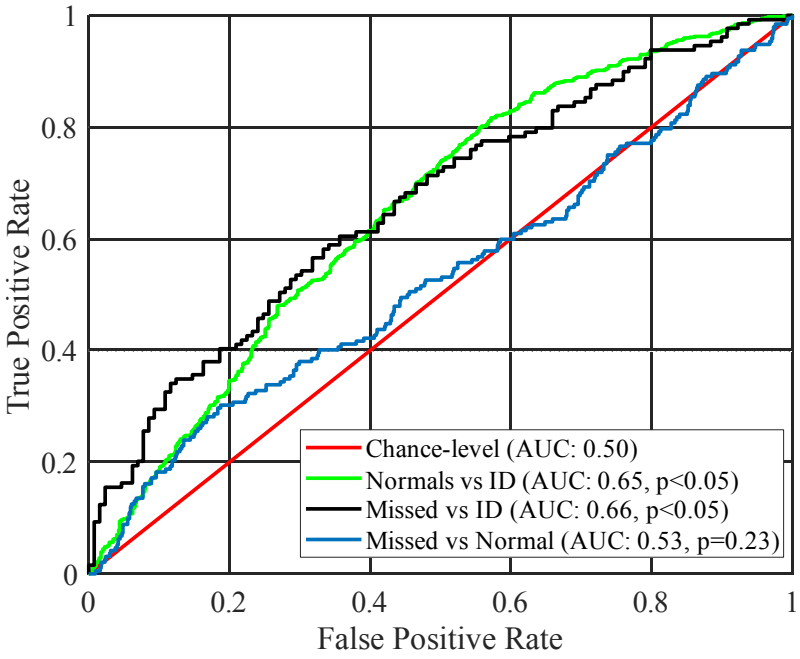


Figure 1- Receiver Operating Characteristics (ROC) curve and Area Under the ROC (AUC) curves for three classifiers, differentiating “Missed” from “ID, “Normals” from “Missed”, and “Normals” from “ID” images. As shown, the performance for “Normal” vs “Missed” classification is at chance-level while other classifier yielded an above-chance level performance. This implies that the global radiomic signature of “Normals” and “Missed” examinations are similar while the signature of both of them is different from the “ID” examinations. This could contribute to the challenging nature of identifying malignancy on “Missed” examination and easier nature of spotting the malignancies on “ID” examination.

Artificial intelligence as a gateway to scientific discovery: Uncovering features in retinal fundus images that allow classification of patient sex via deep learning

Ipek Oruc[1] PhD, Parsa Delavari[1] BSc., Gulcenur Ozturan[1] MD, Ozgur Yilmaz[2] PhD
[1]Department of Ophthalmology and Visual Sciences
[2]Department of Mathematics
University of British Columbia

Rationale

Deep learning (DL) techniques have seen tremendous interest in medical imaging, particularly in the use of convolutional neural networks (CNNs) for the development of automated diagnostic tools. It has recently been shown that patient sex can be predicted based on retinal fundus images using deep learning—a trait that is invisible to the expert human eye in this imaging modality. Here, we investigate the image features that enable the CNNs to categorize sex in retinal fundus images.

Methods

We used the ODIR dataset, a publicly available fundus dataset with 7,000 images, to train a CNN model on sex labels. Images diagnosed with any eye disease or abnormality, as well as images with low quality, were excluded, resulting in 3,146 images (55% male). We fine-tuned a pre-trained Inception-ResNet-v2 model using transfer learning and evaluated it via 5-fold cross-validation. We utilized the Grad-CAM method to obtain attention maps and activation-maximization method to visualize image features that contribute to the model's decision. We specified the area under the receiver operating characteristic curve (AUC) as our model performance metric.

Results

Our model predicts sex with 66.6% AUC, significantly better than chance level ($p<.001$). The model attends to the optic disc, retinal vasculature, and macula to predict the labels. To examine the trained model, we formalized a 3-step discovery pipeline consisting of 1) development of exploratory hypotheses by independent review of visualization outputs, 2) exploratory testing of multiple hypotheses on a small dataset of fundus images ("the sandbox"), and 3) testing of a small number of a priori hypotheses on an independent larger dataset of fundus images ("the test set"). This pipeline resulted in the findings of significantly greater contrast at the border of the optic disc and the peripapillary region in males compared to females; as well as significantly greater horizontal power, likely due to the presence of greater density of vasculature, in the region between the optic disc and the fovea. These findings may suggest anatomical and physiological differences between male and female eyes such as differences in thickness of lamina cribrosa and differences in metabolic activity in the fovea.

Conclusion

We propose a novel methodology, and showcase recent work, to utilize AI in retinal imaging as a promising avenue for scientific discovery. AI applications have shown, implicitly, that the retina is different in males and females and this difference can be detected in fundus images by a trained CNN. In the present work, we show that, interpretation of a deep learning model trained to classify patient sex can lead to the discovery of anatomical and physiological sex differences present in the retina. This approach can be extended to a variety of other avenues of clinical as well as foundational importance.

Interventional X-ray quality assessment using a visibility overshoot index

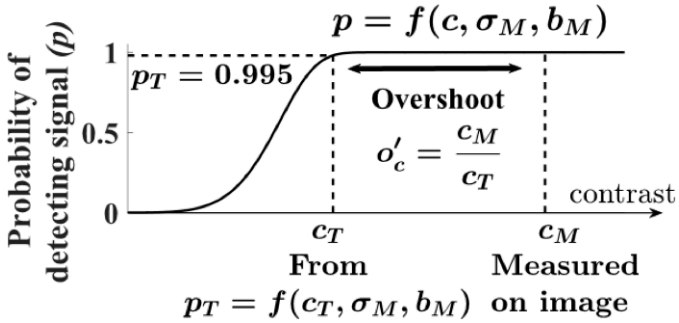
A. Kumcu, MS[1]; L. Platisa, PhD[1]; B. Goossens, PhD[1]; Amber J. Gislason-Lee, PhD[2]; Andrew G. Davies, PhD[2]; Gerard Schouten, PhD[3,4]; Dimitri Buytaert, MS[5]; Klaus Bacher, PhD[5]; W. Philips, PhD[1]

[1]IPI-imec, Ghent University, Belgium; [2]Division of Medical Physics, University of Leeds, UK; [3]Fontys Hogescholen, the Netherlands; [4]Philips Healthcare, the Netherlands; [5]Department of Human Structure and Repair, Ghent University, Belgium

Rationale

X-ray exposure of patients during interventional procedures remains orders of magnitude higher than from modalities such as CT. We propose the *visibility overshoot index* as a new task-based image quality (IQ) measure for dose reduction. We start with two assumptions: clinical structures (signals) must remain visible, and any excess dose above a threshold does not increase the structure's visibility. We can express excess dose in terms of image quality. For example, we know that reducing tube current will reduce dose while increasing image noise. The goal is then to estimate the maximum noise level which maintains signal visibility. Our proposed method measures this threshold and determines the overshoot: the ratio of the threshold to the current IQ level, e.g. the estimated maximum noise level versus the current noise level measured on the image, is the *visibility overshoot index*. The latest automatic dose control (ADC) systems estimate contrast-to-noise ratio by *modeling* physical properties of the signal, X-ray beam, and detector, whereas we *measure* IQ on the acquired patient image, potentially allowing for further dose reduction. The index can also be used to tune image processing algorithms such as denoising. Our task-based approach does not require training like some model observers (MO) and is preferable to assessing fidelity.

Methods



The overshoot index can be expressed in terms of either noise or contrast; the figure illustrates contrast overshoot. The model of signal detection in dynamic noise $p=f(c,\sigma,b)$, derived from staircase experiments we conducted, is the probability p that an observer detects a signal as a function of contrast (c), noise (σ), and background intensity (b). An algorithm we developed segments contrast-filled vessels from patient images and measures IQ parameters in luminance (c_M, σ_M, b_M); these are the model inputs. Above a certain IQ, signal visibility does not increase: detection probability is limited to $0 \le p \le 1$. We find the minimum c_T where the signal remains visible: $p_T = f(c_T,\sigma_M,b_M)$, where p_T=0.995 to constrain the function. The index in terms of contrast is the ratio of measured to threshold contrast: $o_c' = c_M / c_T$. Similarly, $o_\sigma' = \sigma_T / \sigma_M$ is the index in terms of noise. We evaluate the indices against a Channellized Hotelling Observer (d′) for a phantom containing rotationally symmetric objects imaged at dose levels of 6 to 240 nGy/frame. We also compare the indices against clinician's subjective judgments for clinical datasets comprised of two acquisition modes: standard dose, and low dose with denoising.

Results

For the phantom dataset, we found a linear relationship between o_σ' and d′ per signal size, with Pearson and Spearman correlation coefficients between 0.995 and 1. For clinical datasets, o_σ' and o_c' were in concordance with clinicians' judgments.

Conclusions

We present a task-based IQ index for interventional X-ray dose reduction. The proposed index agrees well with a CHO on a phantom dataset, suggesting its potential for accurate IQ

assessment within a dose control loop. Furthermore, the index can be used to evaluate image processing algorithms in a fully automated manner without training or indication of the signals of interest. Additional work is needed to link the index to acquisition parameters. In the future we intend to incorporate signal size in the model.

Phantom data and CHO scores kindly provided by Kenneth Fetterly, Mayo Clinic, Rochester, Minnesota. We thank Robert Hofsink from Philips Healthcare and Dr. Yves Taeymans from Ghent University Hospital for their assistance with this work.

Evaluation of Saliency Models for Clinical Photographs of Disfigured Faces

Haoqi Wang, BS[1,2], Eeshaan Rehani[1], Eloise Jewett[1], Mary Catherine Bordes, BS[2], Krista M. Nicklaus, MSE[1,2], Jun Liu, PhD[2], Gregory P. Reece, MD[2], Summer E. Hanson, MD, PhD[3], Deepti A. Chopra, MD[4], Mia K. Markey, PhD[1,5]

[1]Biomedical Engineering, The University of Texas at Austin, [2]Plastic Surgery, The University of Texas MD Anderson Cancer Center, [3]Section of Plastic and Reconstructive Surgery, [3]University of Chicago Medicine and Biological Sciences, [4]Psychiatry, The University of Texas MD Anderson Cancer Center, [5]Imaging Physics, The University of Texas MD Anderson Cancer Center

Rationale

Head and neck cancer and its treatment can result in facial disfigurement that persists even after reconstructive surgery. Many patients with head and neck cancer express concern about how other people will respond to the changes in their facial appearance. Our long-term goal is to build a saliency model to predict how unaffected individuals perceive facial disfigurement to better prepare patients with facial differences for everyday encounters, e.g., customers in a grocery store. We apply two saliency models to clinical photographs of disfigured faces and use several metrics to evaluate their performance relative to eye tracking data from a human observer study. The purpose this study was to assess the performance of some existing saliency models and to identify opportunities for future model improvements.

Methods

Eye movements were recorded and tracked with a Tobii TX300 Eye tracker (Tobii Technology Inc., Falls Church, VA). 20 lay observers gazed freely on 144 face images of 35 head and neck cancer patients with varying degrees of disfigurement over multiple time points after reconstruction. The fixation data of observers are averaged (examples in figure 1 column 1). Graph-Based Visual Saliency (GBVS) model and the learning-based saliency model of Ren et al. were tested. GBVS works well on natural images but has not been tested on images of disfigured faces. The learning-based saliency model was adapted to use manually selected facial landmarks. We applied similarity metrics to evaluate the performance of saliency models, including Kullback-Leibler divergence, Pearson's correlation coefficient, Area under ROC Curve, etc.

Results

We tested the GBVS and learning-based saliency models on disfigured faces. As expected, the results consistently show that the learning-based model performs better than the GBVS. GBVS detects edges and high contrast regions, but can miss some aspects of the face that are salient to human observers, e.g., the nose. The learning-based saliency model tends to put much weight on the central triangle region of the face and fails to identify disfigurements in the peripheral region of the face as salient (figure 1 row 1). Moreover, despite its emphasis on the central triangle of the face, the learning-based saliency model still underestimates the saliency of midface disfigurements (figure 1 row 2-4). As shown in figure 1, it fails to predict that observers focus on the flap on the left cheek in row 1; the discolored region of the forehead in row 2; the missing left eye in row 3; and the missing nose in row 4.

Conclusions

We present a study that implements two saliency models and evaluates their performances on clinical photographs of disfigured faces. We have shown that the learning-based model is more effective than the bottoms-up algorithm on facial images and we observe the importance of prior weights on the central triangle area of the face. However, the learning-based algorithm underestimated the attention that human observers pay to facial deformities. Off-the-shelf algorithms cannot adequately model human observers viewing photographs of disfigured faces. Improvements in these models are needed to tailor them for our task.

Prior Knowledge of CAD Fallibility Reduces Over-Reliance on the Technology and False Alarms in Mammography

Melina A. Kunar (BSc., PhD) and Derrick G. Watson (BSc., MSc., PhD)

Department of Psychology, The University of Warwick

Rationale

Computer Aided Detection (CAD) has been developed as a useful decision aid in mammography, to highlight areas of potential interest for radiologists to search. Previous research found that when CAD cues are correct, miss errors to find a cancer are reduced (Kunar et al., 2017). However, when CAD cues are incorrect, miss errors and false alarms are greatly increased, showing a strong over-reliance on the technology. The current research investigates whether these over-reliance costs can be mitigated by informing readers of CAD accuracy and performance, prior to the task.

Methods

Forty participants per experiment were recruited from the University of Warwick participant pool. All participants were naïve, non-medically trained observers. The experiments were run on laboratory computers using simulated mammograms, in which the presence of a mass was manipulated. The target prevalence rate varied across experiments (50% in Experiments 1 and 2, 10% in Experiment 3). Prior to each experiment participants were given a training session to familiarise them with mammograms and example masses. Participants were then shown a series of mammograms and asked to search for a mass. CAD cues were used to accurately highlight the mass on 60% of target present trials. CAD cues could also be inaccurate, in which case they highlighted areas with no anomalies. Participants were given information about the accuracy of CAD prior to their search. In Experiment 1, participants were either told that CAD could be beneficial for search, or they were told of its costs (e.g., CAD cues can be inaccurate and cause you to miss a mass). In Experiment 2 and 3, participants were again either told of its benefits or given strong warnings about CAD inaccuracy (e.g. CAD cues are often incorrect, please do not use).

Results

The results showed that purely informing participants about the costs or benefits of CAD did little to alleviate participants over-reliance on the technology. In Experiment 1, miss errors or false alarms did not differ depending on whether participants were told about the costs or benefits of CAD ($F(1, 38) = 0.15$, $p = .70$ and $F(1, 38) = 1.0$, $p = .32$, for miss errors and false alarms, respectively). However, giving readers a strong warning about CAD led to a reduction in false alarms when the CAD cues were inaccurate ($t(38) = 2.29$, $p = 0.03$ and $t(38) = 3.22$, $p = .003$, for Experiments 2 and 3, respectively), while still retaining the benefits of CAD when it detected a cancer.

Conclusions

Giving explicit warnings to readers about the costs of CAD, can help ameliorate over-reliance effects on this technology. This knowledge reduces false alarms, while still retaining the benefits of CAD in terms of cancer detection.

Eye-tracking Differences Between Free Text and Template Radiology Reports

DeAngelo Harris, MD[1], David M. Yousem, MD, MBA[2], Elizabeth A. Krupinski, PhD[1], Mina Motaghi, MD, MPH[2, 3]

[1]Department of Radiology & Imaging Sciences Emory University
[2]Department of Radiology Johns Hopkins Medical Institution
[3]Johns Hopkins University Bloomberg School of Public Health

Rationale

The main communication between radiologists and referring physicians occurs through radiology reports. For more than three decades there have been debates, studies and publications on the quality, format, and content of these reports, all aimed at improving the report value. Conventionally, radiologists report images in free-text format but in 2007 the American College of Radiology published a report encouraging the use of structured templates and in the same year the Radiological Society of North America started an initiative for structured reporting through the RadLex project. Several studies have investigated these two types of report formats from different aspects. Studies have examined clinicians' and radiologists' preferences and expectations, differences in error levels, and quality of reports but there is still debate over the advantages and disadvantages of the structured report format. This study evaluates the difference in the time radiologists look at the image when using free text vs. structured templates. We hypothesized that the total amount of time spent viewing the image is diminished when a structured template report is employed compared to free texting and/or the total time to create a dictation on a report will be prolonged with structured template reporting.

Methods

A cohort of neuroradiologists and 2nd/3rd year residents served as subjects. Each viewed a series of 5 de-identified brain images and compiled a report, one using a free text form and the other a structured template. Sessions were counterbalanced with at least 3 weeks between sessions. Eye-tracking was done using a Tobii system. Demographic data were collected including age, sex, year of residency, and current report format preference.

Results

Data to date reveal differences between individuals in terms of preferred report format. The eye-tracking data are still being evaluated, but there appear to be differences between time spent viewing the images vs the report for free format versus structured templates as well as the number of shifts in attention between the two displays.

Conclusions

There may be significant differences in the amount of time spent viewing the report display vs the image display when using free format vs structured templates for reporting. This may have implications for discovering underlying sources of errors in reporting as well as reporting efficiencies. Further studies are required with more and different types of errors.

Localization ROC Analysis Revisited

Yulei Jiang, Ph.D.
Department of Radiology, The University of Chicago

Rationale

Receiver operating characteristic (ROC) analysis is commonly applied to evaluate detection performance where an observer does not know a priori whether a lesion is present nor the location of the lesion if one is present. A common observer error, recognizable by lesion localization, is to miss a ground-truth lesion and to identify a pseudo lesion that is a false positive. But ROC analysis of this type of observer error has been unsatisfactory.

Methods

We show that ROC analysis cannot naturally analyze this type of observer error because fundamentally this is a missing data problem: the observer report on the ground-truth lesion is missing. Rather than a single ROC curve that characterizes the detection performance, we at best obtain two ROC curves that bracket this performance. The upper limit ROC curve is the socalled case-based ROC curve wherein the observer error in lesion localization is ignored. This, in effect, assumes that the ground-truth lesion will be detected eventually during subsequent clinical work-up of the case even though it was missed initially. The lower limit ROC curve is the localization ROC (LROC) curve wherein cases that contain this type of observer error are removed from the ROC analysis and analyzed separately under the assumption that the missed ground-truth lesion will not be detected (Fig. 1).

Results

This new LROC curve is identical to Starr's LROC curve, despite that Starr's original proposal applies to only a narrowly defined detection task and assumes certain rational observer behavior whereas this new LROC curve applies to detection tasks in general and makes no assumption on observer behavior. The new LROC curve can be estimated via maximum likelihood and, unlike Starr's LROC curve, cannot be predicted from a related ROC curve. Simulations and example reader datasets provided validation evidence.

Conclusion

The LROC analysis is a meaningful alternative for characterization of detection performance that includes observer errors when a ground-truth lesion is missed and a false-positive lesion is identified.

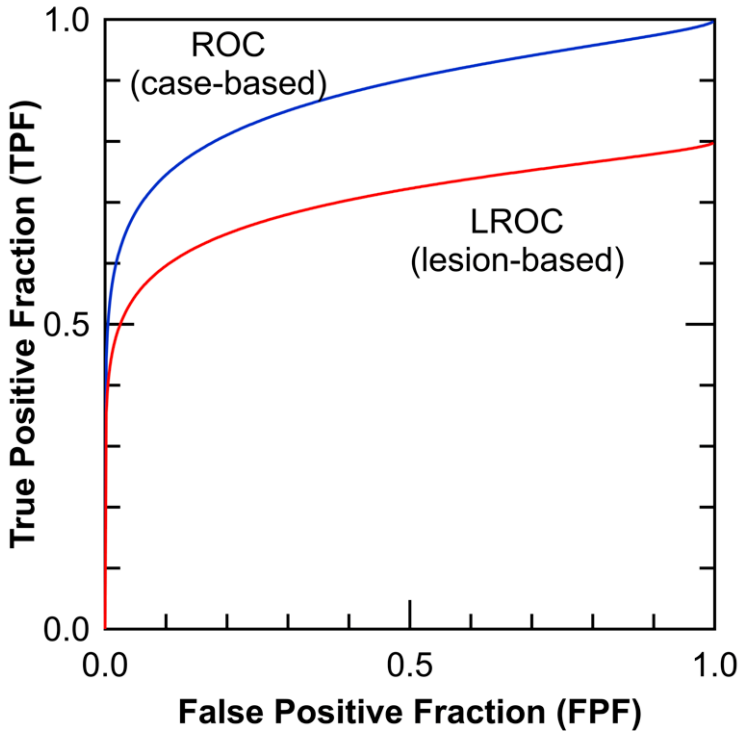


Figure 1. Detection performance is characterized by two limiting ROC curves: The case-based ROC curve ignores observer errors when a ground-truth lesion is missed but a pseudo (falsepositive) lesion is identified and constitutes the upper limit by, in effect, assuming missed ground-truth lesions will be detected eventually during subsequent clinical work-up of the case. The LROC curve assumes missed ground-truth lesions will not be detected and constitutes the lower limit. Characterization of detection performance cannot be more precise than the above without additional knowledge or assumption of detection outcomes of missed lesions.

Assessing Satisfaction of Search in Virtual Breast Images for Experts and Novices

Stephen H. Adamo[1], Miguel Lago[2], Bruno Barufaldi[3], & Joseph Schmidt[1]
University of Central Florida, University of California Santa Barbara, University of Pennsylvania

Rationale

While there are many reasons for search errors within breast imaging (Gandomkar & Mello-Thoms, 2019) what is unknown whether Satisfaction of Search (SOS) occurs. SOS is a phenomenon where searchers are more likely to miss a lesion/target after detecting a first lesion/ target. Cognitive research suggests that detection of one target improves detection for similar compared to different targets (Biggs et al., 2015). Our goal is to investigate SOS in breast imaging and whether the detection of a specific lesion (e.g., a mass) impacts detection differently for a similar lesion (e.g., another mass) compared to a different lesion (e.g., a calcification) and expertise mitigates the SOS effect.

Methods

Thirty-four radiologists with breast imaging experience were recruited from the Radiological Society of North America (RSNA) conference and 42 novices were recruited from the University of Central Florida. The OpenVCT framework was used to simulate the breast anatomy of patients. The digital breast tomosynthesis (DBT) reconstructions and synthetic images were simulated using simple back projection with a commercially available software library (Briona, version 7.12, see Barufaldi, 2018). Up to two simulated lesions were inserted at random x and y coordinates with 7 containing no lesions, 30 containing one, and 15 containing two (5 dual-mass, 5 dual-calcifications, and 5 mass and calcification images). We followed the methods of Adamo et al., (2020) and images were created as near-identical triads with two, single-images paired with every dual image. *This design keeps everything that perceptually impacts detection (e.g., breast density) constant, and only the impact that a first lesion has on a second lesion was quantified. Virtual breast images allowed us to create three identical copies of the same breast image with only abnormalities added or removed.*

Results

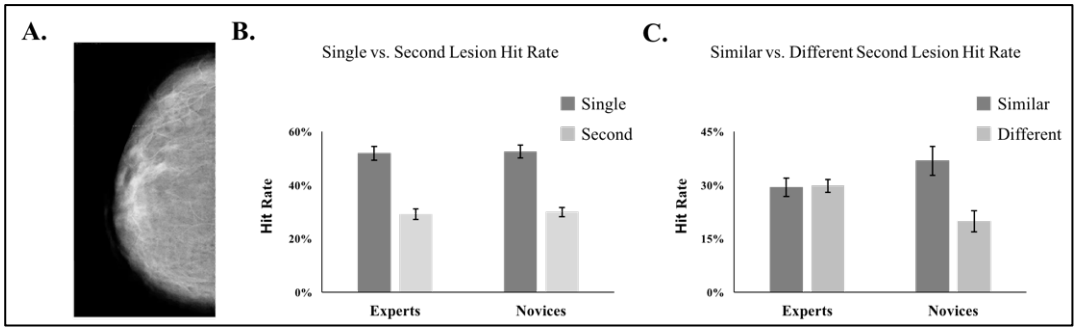


Figure: A: Breast phantom image. B: Graph of single and second lesion hit rate (i.e., to determine whether there was an SOS effect). C: Graph of same and second lesion hit rates when they were similar and dissimilar to a detected first lesion (i.e., to determine whether there was a similarity effect).

<u>*Main effects:*</u> A 2x2 between subjects' ANOVA of single- vs. second-lesion accuracy and lesion similarity between experts and novices was performed. The main effect of single vs. second was significant ($F(1,73)=129.5$, $p<.001$), with a 21.0% SOS effect. The main effect of similarity was also significant ($F(173)=34.3$, $p<.001$). The between subjects effect of expertise was non-significant ($F(173)=0.2$, $p=.88$). The interaction of similarity by expertise ($F(1,73)=6.54$, $p=.01$) was significant, and single vs. second by similarity was significant ($F(1,73)=8.32$, $p=.005$). Follow-up t-tests

determined that experts (M=29.8%) were more accurate than novices (M=18.9%; $t(74)=2.04$, $p=.045$) when lesions were different, but were less accurate when lesions were similar (experts: M=29.4%, novices: M=36.7%; $t(74)=2.43$, $p=.018$).

Conclusions

In this study, we demonstrated that SOS errors are a viable cause for lesion misses and expertise impacts second lesion misses after first lesion hits. Future research will need to investigate potential differences in eye-movement strategies and the role familiarity (e.g., long-term vs. short-term memory) of lesion/targets to determine why second-target error rates differed by expertise.

Optimizing the set of pairs of radiologists that double read screening mammograms

Jessie J.J. Gommers, MSc[1] Dr. Craig K. Abbey, PhD[2] Dr. Fredrik Strand, PhD[3,4] Prof. Sian Taylor-Phillips, PhD[5] Marthe Larsen, MSc[6] Prof. Solveig Hofvind, PhD[6,7] Prof. Mireille J.M. Broeders, PhD[8,9] Prof. Ioannis Sechopoulos, PhD[1,8,10]

[1]Department of medical imaging, Radboudumc, Nijmegen, The Netherlands [2]Department of psychological and brain sciences, University of California, Santa Barbara, United States [3]Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden [4]Breast Radiology, Karolinska University Hospital, Stockholm, Sweden [5]Warwick Medical School, University of Warwick, Coventry, United Kingdom [6]Section of Breast Cancer Screening, Cancer Registry of Norway, Oslo, Norway [7]Department of Health and Care Sciences, UiT The Arctic University of Norway, Tromsø, Norway [8]Dutch expert center for screening (LRCB), Nijmegen, The Netherlands [9]Department for Health Evidence, Radboudumc, Nijmegen, The Netherlands [10]Technical Medicine Centre, University of Twente, Enschede, The Netherlands

Rationale

A potential solution to reduce radiologist errors may be to optimize the double reading of screening mammograms. The pair composed of the two best radiologists of the whole program would yield the best outcome, but caseloads need to be divided evenly. We aim to investigate how radiologist performance characteristics can be leveraged to determine the optimal set of pairs of radiologists for the double reading of screening mammograms.

Methods

We retrospectively analyzed three datasets of screening examination outcomes of women who underwent screening mammography in Sweden, the United Kingdom (UK), and Norway. The screening examinations were double read and any examination that was flagged by either radiologist was classified as abnormal. Cancer detection rates (CDR) and abnormal interpretation rates (AIR) were evaluated for individuals and pairs of radiologists. The individuals were divided into four performance categories involving CDR and AIR: high CDR & low AIR (HL), high CDR & AIR (HH), low CDR & AIR (LL), or low CDR & high AIR (LH). Based on the four different types of individual readers, 10 different types of pairs exist, each having their own average CDR and AIR. Random pair performance, for which any pair was equally likely, was compared to the performance of specific pairing strategies. Bootstrap resampling was used to obtain 95% confidence intervals.

Results

For all three datasets, the CDRs for the specific pairings were not statistically significantly different from the CDRs for the random pairings. The Swedish and UK datasets did show a significant similar pattern for AIR: compared to random pairings, pairing strategies with opposite AIR radiologists resulted in a significant AIR reduction of 3.4% and 2.9%, for the Swedish and UK dataset, respectively. For the Swedish dataset, the pairing strategy with fully opposite performance radiologists also resulted in a significant 10.3% AIR reduction when compared to random pairing. The Norwegian dataset, however, showed a different pattern, with no significant differences between the AIRs of specific pairings and the random pairing.

Conclusions

Pairing radiologists based on their performance characteristics, as opposed to randomly, may improve grouped screening performance. However, our data showed contradicting patterns for the different pairing strategies. Further analyses with more datasets from different screening settings have to be done to investigate how radiologists performance characteristics can potentially be used for the pairing of radiologists. A limitation of our datasets was that each examination was read by only two radiologists, so only existing pairings could be analyzed. Our plan is to start modelling radiologist AIR decisions, so that we can explore all possible pairs.

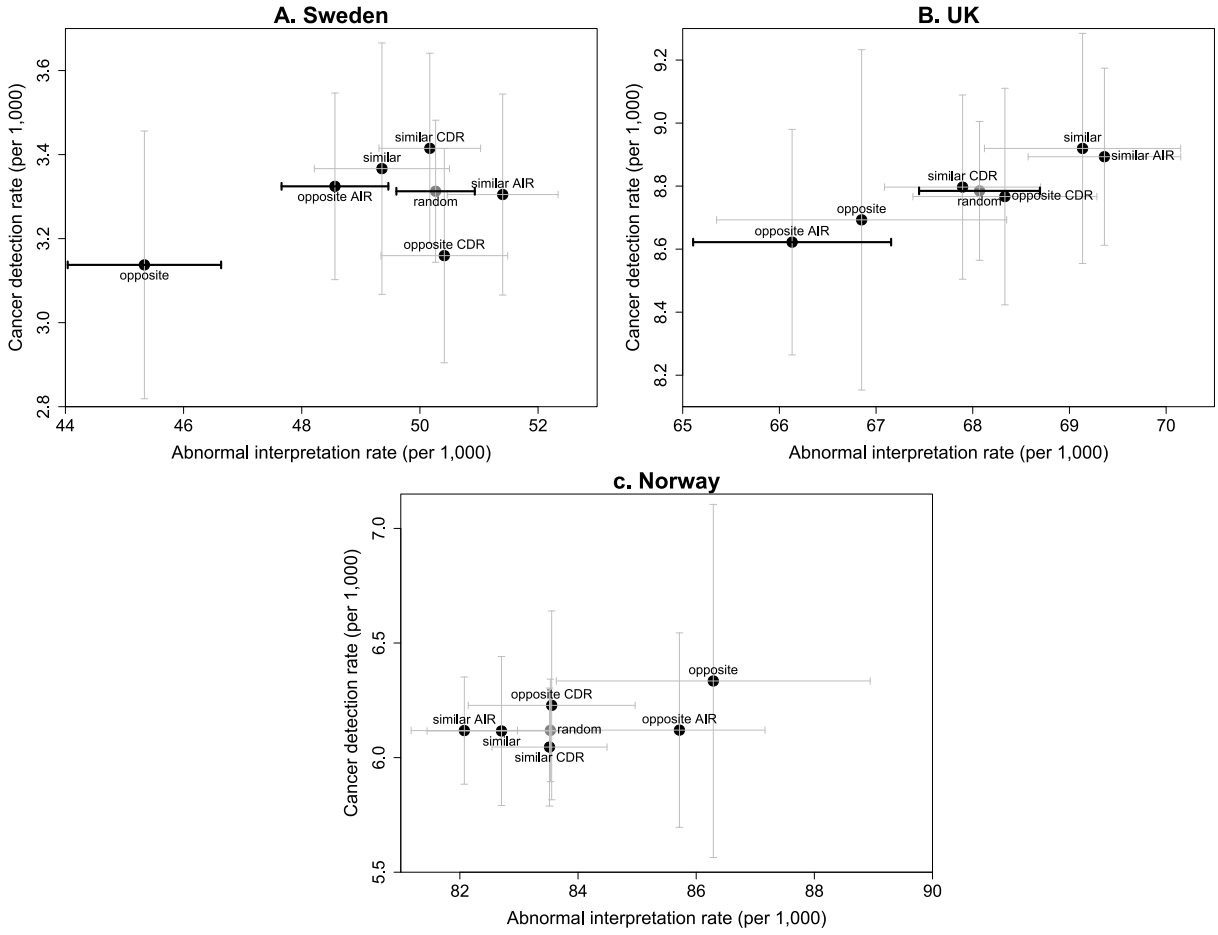


Figure – Screening performance for the different pairing strategies The dots represent the average screening performance for the random pairing (red) and the specific pairing strategies (black). Error bars are 95% confidence intervals. The bold error bars demonstrate pairings with performance that is significantly different from the random pairing. Please note that the axes are different, due to the differences in CDR and AIR for the three datasets. AIR, Abnormal Interpretation Rate; CDR, Cancer Detection Rate.

Modeling Search-time Behavior in a Satisfaction of Search (SOS) Experimental Framework: The Role of Context and Experience.

Nelson A. Roque, PhD[1], Stephen Adamo, PhD[1] Miguel A. Lago[2], Bruno Barufaldi[3],
[1]University of Central Florida, [2] University of California Santa Barbara, [3] University of Pennsylvania

Rationale

Faster responding is expected as people practice or learn a system or interface (e.g., a medical image system for breast images). However, it is unclear how learning is affected by expertise in breast cancer detection when searching for multiple abnormalities. This study, leveraging a generalized linear mixed modeling (GLMM) framework, intended to model search time as a function of stimuli evaluation conditions (trial number, number of targets) and expertise (novices vs. residents/fellows/radiologists with breast imaging experience) to better understand how context and experience can shape search performance.

Methods

Seventy-five participants (42 undergraduate students; 33 radiologists/residents/fellows with breast imaging experience), searched up to two masses and/or calcifications in a multi-abnormality virtual digital breast tomosynthesis (DBTs). These DBTs were obtained with the OpenVCT imaging pipeline and included Coopers' ligaments, skin, adipose, and glandular tissue compartments. The DBTs were 700ml with 6.33cm ML-compressed thickness, 100-micrometer voxel size, various simulated parenchymal patterns, and 15-25% of glandular tissue. The DBTs were created with abnormalities on the center 2D slice, which was then used as the image in the experiment. Participants completed 52 trials of the search task, i.e., 30 single target trials, 15 dual target trials, and 7 target absent trials. This study leverages mixed-effect modeling (using R package lme4) to account for inter-observer differences (i.e., random intercept per participant) across the entire dataset, while simultaneously exploring effects of search time in study (i.e., practice effect). An unconditional means model is the baseline used to compare subsequent models with additional fixed effects (e.g., display type; dual-target, single-target trials).

Results

A model containing fixed effects for trial number (as a proxy for practice / learning effects), number of targets, and whether or not the participant has breast imaging experience was the best performing model, when compared with the unconditional means model, and each prior model step (p's $< .001$ for all model comparisons). First, the marginal R2 (0.079) was markedly different from the conditional R2 (0.428), suggesting that a substantial amount of variation in search time is explained by inter-observer intercept differences. In general, participants responded significantly more quickly for each trial ($\beta = -0.13$, $p < .001$), for each unit increase in number of abnormalities ($\beta = -2.98$, $p < .001$), and if they had breast imaging experience ($\beta = -10.19$, $p < .001$). In other words, experience matters – both in terms of individual differences (e.g., breast imaging experience), but also within the experiment (i.e., responding more quickly over time) – and these distinct but related sources of variation should be carefully considered in a chosen analytic framework.

Conclusions

These data serve as initial validation of leveraging a generalized linear mixed modeling (GLMM) framework to simultaneously account for sources of variation at different levels of a medical image search experiment. Given that substantial variation was explained by between person-differences in search time (i.e., participant intercepts), future work may evaluate what predicts these differences, beyond expertise class alone (e.g., years of experience, conscientiousness, level of fatigue, time of day, etc.).

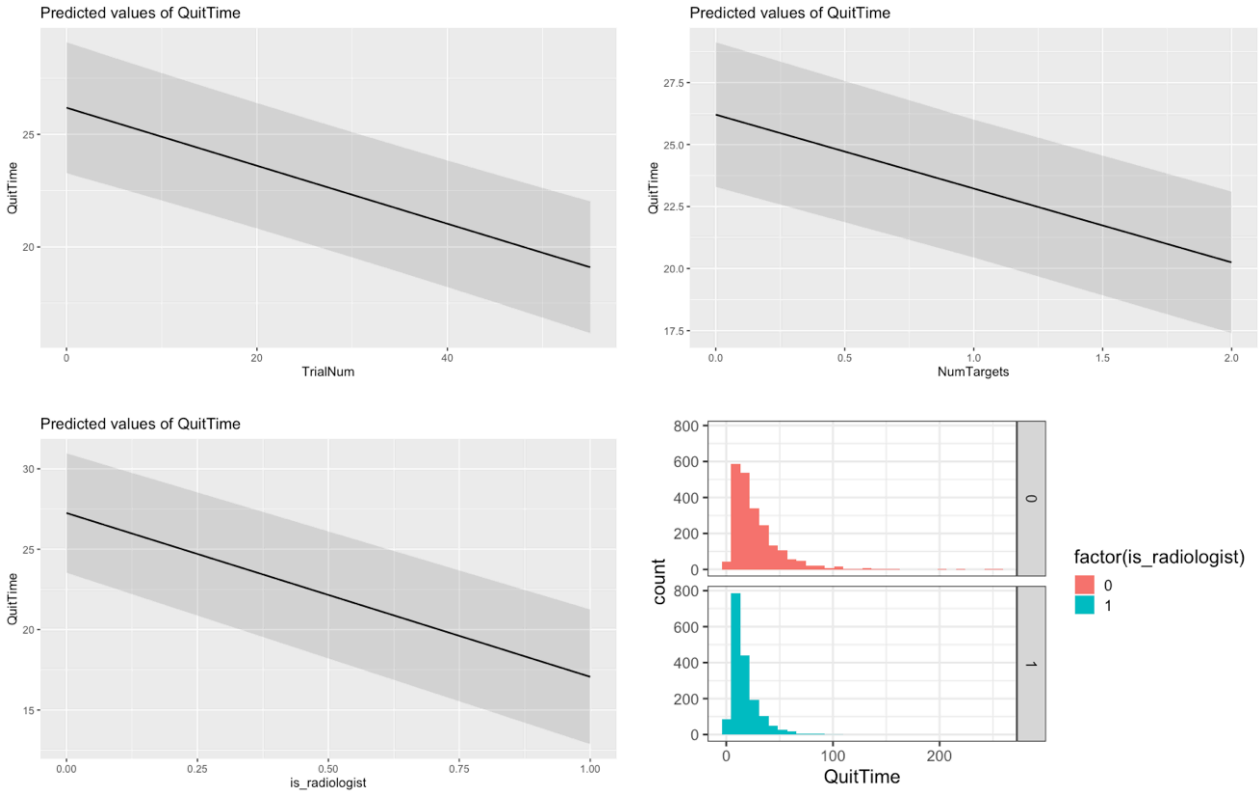


Figure 1. (Top, left) Predicted values of search time as a function of trial number. Interpretation, participants are quitting more quickly towards the end of the experiment. (Top, right) Predicted values of search time as a function of number of targets. Interpretation, participants are quitting more quickly as a function of more targets. (Bottom, left) Predicted values of search time as a function of expertise (0 = undergraduate students; 1 = radiologist). Interpretation, radiologist participants are quitting more quickly in general. (Bottom, right) Histogram of search time by participant group (0 = undergraduate students; 1 = radiologist).