CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

1

Medical image perception

EHSAN SAMEI AND ELIZABETH KRUPINSKI

1.1 PROMINENCE OF MEDICAL IMAGE PERCEPTION IN MEDICINE

Medical images form a core portion of all the information a clinician utilizes to render diagnostic and treatment decisions while a patient is under his/her care. As such, medical imaging is a major feature of modern medical care. An important requirement in using medical images is to understand what an image indicates; there is therefore a need to perceive (i.e. interpret) medical images and an associated need to have physicians subspecialized in medical image interpretation. The goal of this chapter is to provide a broad picture of the importance of medical image perception from a general healthcare perspective.

Medical imaging has been primarily ascribed to the subspecialty of radiology, with about a billion radiological imaging exams performed worldwide every year. The images include many types of examinations – single projection X-rays used in musculoskeletal, chest, and mammography applications; dynamic X-ray exams such as fluoroscopy, three-dimensional computed tomography (CT), and magnetic resonance (MR) exams; nuclear medicine emission images; and ultrasound. With the advent of digital imaging and multi-detector CT, the type and number of radiology examinations have been changing as well. The range of image types is also expanding rapidly with new modalities such as tomosynthesis and molecular imaging, which is being investigated for numerous applications, from identifying lesion margins during surgical removal to identifying cancer cells in the blood. Imaging technologies are extremely varied. Medical images can be grayscale or color, high-resolution or low-resolution, hardcopy or softcopy, uncompressed or compressed (lossy or lossless), acquired with everything from sophisticated dedicated imaging devices to off-the-shelf digital cameras.

While imaging is the central technology behind the subspecialty of radiology, during the past several years, imaging has also expanded beyond radiology to embrace other subspecialties including cardiology, radiation oncology, pathology, and ophthalmology, to name a few. Study of pathological specimens used to be limited to glass slide specimen "images" rendered by the microscope for the pathologist to view. With the advent of digital slide scanners in recent years, however, virtual slides are becoming more prevalent not only in telepathology applications but in everyday reading (Weinstein, 2001). In many medical schools and pathology residency programs, students are no longer required to purchase a microscope and a box of glass specimen slides. Students now learn from a CD with directories of virtual slides to view as softcopy images. Ophthalmology has relied on images for years (mainly as 35 mm film prints or slides) for evaluating such conditions as diabetic retinopathy. With the advent of digital images and high-performance color displays, screening raters are increasingly using softcopy images. Telemedicine has opened up an entirely new area in which medical images are being acquired, transferred, and stored to diagnose and treat patients (Krupinski, 2002). Specialties such as teledermatology, teleophthalmology, telewound/burn care, and telepodiatry are all using images on a regular basis for store-and-forward telemedicine applications. Real-time applications such as telepsychiatry, teleneurology and telerheumatology similarly rely on video images for diagnostic and treatment decisions.

One way to demonstrate the pervasiveness of medical imaging is to examine the amount of money spent each year on healthcare and then portion out the amount devoted to medical imaging (Beam, 2006). Relying on 2004 data from the Centers for Medicare and Medicaid Services (CMS), approximately 16% of the gross domestic product (GDP), or $1.6 trillion, is allotted to national healthcare expenditures (http://www.cms.hhs.gov/home/rsds.asp). Medicare expenditures represent 17% of national healthcare expenditures, of which Part B (43%) accounts for the non-facility or physician-related expenditures. Approximately 8% of Part B (or nearly $10 billion) constitutes physician-based imaging procedures. Imaging also accounts for over 40% of all hospital procedures reported in the discharge report according to the Agency for Healthcare Research and Quality (AHRQ) (http://www.ahrq.gov/data/hcup/). Based on Medicaid Part B spending, one can conservatively assume that imaging procedures comprise only 8% of non-Medicaid Part B health spending. Therefore, medical imaging in the USA is estimated to amount to $56 billion ($10 billion/17%/43%), or 0.5% of GDP.

With the pervasiveness of imaging in modern medicine, there has been significant attention and interest in the technology of imaging operations, ranging from hardware features to software functionalities. What is less appreciated is the perceptual act underlying the interpretation of these images (Manning, 2005). In order to impact patient care, an image must be *perceived and interpreted* (i.e. understood in the context of patient care) (Figure 1.1). If one assumes each of the one billion imaging examinations performed worldwide annually involves an average of four individual images per exam, one could compute that on the average, 120 medical image perception events take place every second! This astounding frequency speaks further of the pervasiveness of medical image perception in healthcare enterprise.

The Handbook of Medical Image Perception and Techniques, ed. Ehsan Samei and Elizabeth Krupinski. Published by Cambridge University Press.
© Cambridge University Press 2010.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

Figure 1.1 As a fundamentally visual discipline, medical imaging requires psychophysical interpretation of the images to draw "meaning" from the imaging information and understand their clinical relevance.

Figure 1.2 The detection of a subtle abnormality is somewhat similar in difficulty to identifying the dog in a popular visual demonstration.

The need for interpretation of medical images comes from the fact that medical images are not self-explanatory. In the popular culture, "a picture is worth a thousand words," a phrase that reflects the power and utility of images. Ironically however, the interpretation of a medical image involves summarizing a multi-dimensional image into a few words because medical images *by themselves* do not deliver the certainty that they promise (Figure 1.2). This uncertainty, which necessitates interpretation, stems from the nature of medical imaging. Imaging is ultimately a visual discipline, impacted by psychophysical processes involved in the interpretation of images. For example, medical images can contain anatomical structures that can camouflage a feature of clinical interest that is not prevalent (in the case of screening). This uncertainty impacts the psychology of interpretation. Added to this complexity are notable variations from case to case and a multiplicity of compounding abnormalities and related factors that the interpreter needs to be mindful of.

There are clearly a significant number of images being viewed and interpreted by clinicians today in a variety of clinical specialties. As such, diagnostic accuracy cannot be defined independently of the interpretation, and any limitations or suboptimality in terms of how the images are used can significantly influence the diagnostic and therapeutic clinical decisions that they enable. Given a one-to-one link between an image and its interpretation, imaging technology alone can offer little in terms of patient care if the image is misinterpreted. The complexities of image interpretation can lead to interpretation errors and clinicians do make mistakes in the interpretation of image data (Berlin, 2005, 2007). Estimates in radiology alone suggest that in some areas there may be up to a 30% miss rate and an equally high false-positive rate. Errors can also occur in the recognition of an abnormality (e.g. whether a lesion is benign or malignant). Such errors can have a significant impact on patient care due to delays or misdiagnoses. What is less well appreciated is the prominent contribution of the inherent limitations of human perception to these errors. Image perception is the most prominent yet least appreciated source of error in diagnostic imaging. The prominence of imaging reading errors in malpractice litigation is an example of this ignorance.

The likelihood of error in the interpretation of images emphasizes the need to understand how the clinician interacts with the information in an image during interpretation. Such an understanding enables us to determine how we can further improve decision-making. That brings us to the science of medical image perception. Error is one reason to study medical image perception.

1.2 THE SCIENCE OF MEDICAL IMAGE PERCEPTION

First and foremost, it is important to understand the nature and causes of interpretation error. For that objective, one needs to distinguish between visual errors (estimated to amount to about 55% of the errors) because the clinician does an incomplete search of the image data (Giger, 1988); and cognitive errors (45%), where an abnormality is recognized but the clinician makes a decision-making error in calling the case negative (Kundel, 1978). Visual errors are further subdivided into errors where the clinician fails to look at the territory of the lesion (30%) (Kundel, 1975, 1978), and those when he/she does not fixate on the territory for an adequate amount of time to extract the lesion's relevant features (25%) (Carmody, 1980).

Contributing to interpretation errors are a host of psychophysical processes. Camouflaging of the abnormality by normal body features (so called anatomical noise) is one of the main contributors to interpretation error. Masking of subtle lesions by normal anatomical structure is estimated to affect lesion detection threshold by an order of magnitude (Samei, 1997). The visual search process, necessitated by the limited angular extent of the high-fidelity foveal vision of the human eye, is another important contribution to image interpretation. Preceded by a global impression or gist, a visual search of an image involves moving the eye around the image scene to closely examine the image details (Nodine, 1987). Studies on visual search have highlighted the prominent role of peripheral vision during the interpretation, where there is an interplay between foveal and peripheral vision as the observer scans the scene (Kundel, 1975). As a result, there are characteristic dwell times associated with correct and incorrect decisions that are influenced by the task and idiosyncratic observer search patterns (Kundel, 1989). Satisfaction of search – once an abnormal pattern is recognized, it takes additional

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

diligence on the part of the clinician to look for other possible abnormalities within an image – is yet another contributing factor to errors (Berbaum, 1989; Smith, 1967; Tuddenham, 1962, 1963). Studies have explored the impact of expertise and prior knowledge in that behavior.

Image quality is yet another topic of interest. While intuitively recognized, image quality has been more elusive than image interpretation to characterize in such a way that it would directly relate to diagnostic accuracy (or its converse, diagnostic error). In that regard, it is important to understand how best to assess image quality and its impact on perception in order to optimize it and minimize error (Krupinski, 2008). Studies have focused on the impact of image acquisition, imaging hardware, image processing, image display, and reading environment on image quality and diagnostic accuracy.

Ergonomic aspects of interpreting medical images also play a role in the perception process. There is a need to understand the impact of ergonomic and presentation factors to minimize error (Krupinski, 2007), including determination of the causes of fatigue and how they can be minimized, the contribution of fatigue to error, the environmental distractions, the impact of the viewing interface, especially with softcopy images, and the impact of the color tint of the image.

Though we hope and aim for consistent and correct clinical decisions with every case, that aim is hard to achieve. The likelihood of two clinicians rendering two different interpretations of the same image is unsettlingly high and the expertise of the clinician plays an important role in this problem. Medical expertise is the ability to *efficiently* use contextual medical knowledge to make accurate and consistent diagnoses. Medical imaging expertise further involves perceptual and cognitive analysis of image features and manifests itself in a rich structured knowledge of normalcy and "perturbations" from the normal, an efficient hypothesis-driven search strategy, and an ability to generalize visual findings to idealized patterns. Achieving such expertise requires talent further honed by motivated effortful study, preferably supervised, and dedicated work, where accuracy is roughly proportional to the logarithm of the number of cases read annually (Nodine, 2000). Topics of interest in this line of investigation include the impact of the clinician's experience, age, and visual acuity on accuracy, toward better training and utilization of medical imaging clinicians.

Considering the impact of image perception on diagnostic accuracy, it is often necessary to test various imaging technologies and methods in terms of the associated impact on image perception. Such studies require the use of experienced clinicians, which is an expensive undertaking. Thus, there is a great need for accurate computational programs that can model visual perception and predict human performance. A host of such perceptual models have been developed, including the ideal human observer model, non-prewhitening models, channelized models, and visual discrimination models (Abbey, 2000). These models naturally require a reasonably accurate understanding of the image interpretation process. As our knowledge of the process is limited, so is the accuracy of these models. As such, their use often requires certain assumptions, verifications of their accuracy and relevance in pilot experiments, and certain calibrations, e.g. adding internal noise to make the model predictions fit human results. Nonetheless, these models have demonstrated valuable, though limited, utility in many applications, and their advancement continues to shed light on the image interpretation process.

By and large, image interpretation is currently a human task. However, increasingly, artificial intelligence tools are being used to aid in interpretation or to replace the radiologist altogether. The most common technology currently used is computer aided diagnosis (CAD), computer algorithms that examine the image content for certain abnormal features of clinical interest and then prompt the clinician for a closer examination of those features (Doi, 2007). CAD is becoming an important tool for interpreting medical images, considering the exponential growth of imaging and the shortage of specialized expertise. There is currently a need to understand the impact of CAD on diagnosis by investigating issues such as how best to integrate the human and the machine in such a way that the strength of both can be fully utilized towards improved diagnosis. For example, an experienced clinician might ignore the CAD prompts or be distracted by them if the system indicates too many false-positives. On the other hand, an inexperienced clinician might overly depend on CAD, initiating unnecessary follow-up procedures or dismissing an abnormality that might not have been picked up by the CAD algorithm. Such patterns might also change over time as a clinician gets used to a system, and such "getting used to" might not necessarily lead to improved diagnosis or efficiency. Thus, there is a need to understand the impact of CAD on the clinician's psychology, expertise, efficiency, and specialization paradigms.

Fundamental to the discussion above is the need to measure diagnostic accuracy itself (Metz, 2006; Wagner, 2007). There are a number of measures of performance such as fraction correct, sensitivity, and specificity. However, such simple measures do not adequately reflect accuracy as they can be dependent on disease prevalence or the criteria applied by the clinician, e.g. a clinician who calls all cases abnormal will have a perfect sensitivity but poor specificity, and vice versa. Seeking an overall performance measure independent of disease prevalence and criterion, receiver operating characteristic (ROC) analysis has emerged as the current gold standard for measuring diagnostic accuracy. However, ROC analysis has a number of limitations, including being limited primarily to single tasks, nonbinary confidence ratings, and location-independent decisions. In recent years, a number of advances of the ROC methodology have been developed, a welcome expansion which has shown continued progress.

1.3 WHY A CLINICIAN SHOULD CARE ABOUT MEDICAL IMAGE PERCEPTION

Medical image perception is a mature science that continues to be advanced by expert scientists. When over-specialization causes specialized "territories" to be left to the experts, one may ask why a clinician who interprets medical images needs to care about medical image perception. Needless to say, no one expects a clinician to also be a medical perception scientist. However, some knowledge of perception issues and concerns can provide

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

vital advantages for the clinician who interprets medical images. Those advantages can be grouped into five categories.

1. Patient care-related: Understanding perceptual issues could help a clinician to improve his/her performance. Knowledge of key perceptual factors such as satisfaction of search, the relevance of prolonged dwell time, search strategies, and psychological impacts of CAD can affect the way he/she interprets medical images. Awareness of these issues enforces a greater care about the way the images are created, a greater appreciation for image quality and its effect on accuracy and efficiency, an appreciation for the influence of fatigue and the proper ergonomic design of the working environment, and higher confidence in the use of new display technologies.
2. Science-related: Being better informed about key perceptual factors enables a more proper design of projects involving medical images, develops an ability to better answer perceptual questions that inevitably arise in the review of imaging-related papers and grant applications, and increases proficiency in the reviewing of such papers and grants.
3. Teaching and learning-related: Knowledge of perceptual factors can help clinicians better communicate their expertise to trainees and help clinicians hone their perceptual skills.
4. Consumer-related: Understanding the importance of perceptual factors enables a clinician to be a better shopper of medical image-related products and services. For example, he/she will be more mindful of the image quality performance of acquisition and display devices, and the importance of the graphical user interface of picture archiving and communication system workstations.
5. Profession-related: Awareness of image perception issues enables a clinician to better educate patients, other medical professionals, and the public about the statistical nature of medical image interpretation, and to play a more effective role in related malpractice litigations.

1.4 ABOUT THIS BOOK

As outlined above, medical image perception is a frequent clinical task and a notable component of modern medicine. With perceptual error as one of the major sources of medical decision errors, our knowledge of perceptual issues gives us resources to minimize these errors and to educate future medical imaging clinicians and scientists. This book aims to provide a comprehensive reflection of medical perception concepts and issues within a single volume. Chapters in this text deal with a variety of perceptual issues in detail.

The first part of the book offers chapters by four prominent scientists, reflecting on historical developments of the field and its theoretical foundations. This part includes some reflections of the late Robert Wagner, the legendary perception scientist whose work and impact has been paramount in shaping the field as we know it today. The second part of the book includes six chapters discussing the science of medical image perception. Main topics include visual and cognitive factors, satisfaction of search, and the role of expertise. This part concludes with the perceptual relevance of image quality and reflections on the limitations of the human visual system. Part three focuses on perception metrology, with chapters on the logistics of designing perception experiments, and ROC methodology and its variants. This part ends with discussion of perceptual observer models and their implementation. Part four focuses on decision support and CAD, with topics ranging from the design of CAD studies to perceptual factors associated with the use of CAD in interpreting chest, breast, and volumetric images.

The last major part of the book offers six additional chapters about specific optimization considerations from a perceptual standpoint. Applications include radiography, CT, mammography, image processing, and image display. This part further offers a perspective on ergonomic design of workplaces for radiologists. The book ends with an epilogue outlining future possible directions for medical image perception science.

REFERENCES

Abbey, C.K., Bochud, F.O. (2000). Modeling visual detection tasks in correlated image noise with linear model observers. In Beutel, J., Van Metter, R., Kundel, H. (eds). *Handbook of Medical Imaging, Vol. 1: Physics and Psychophysics*. Bellingham, WA: SPIE Press, pp. 655–682.

Beam, C.A., Krupinski, E.A., Kundel, H.L., Sickles, E.A., Wagner, R.F. (2006). The place of medical image perception in 21st-century health care. *JACR*, **3**, 409–412.

Berbaum, K.S., Franken E.A., Dorfman, D.D., *et al.* (1989). Satisfaction of search in diagnostic radiology. *Invest Radiol*, **25**, 133–140.

Berlin, L. (2005). Errors of omission. *AJR*, **185**, 1416–1421.

Berlin, L. (2007). Accuracy of diagnostic procedures: has it improved over the past five decades? *AJR*, **188**, 1173–1178.

Carmody, D.P., Nodine, C.F., Kundel, H.L. (1980). An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*, **9**, 339–344.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imag & Graphics*, **31**, 198–211.

Giger, M.S., Doi, K., MacMahon, H. (1988). Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. *Med Phys*, **15**, 158–166.

Krupinski, E.A., Jiang Y. (2008). Evaluation of medical imaging systems. *Med Phys*, **35**, 645–659.

Krupinski, E.A., Kallergi, M. (2007). Choosing a radiology workstation: technical and clinical considerations. *Radiol*, **242**, 671–682.

Krupinski, E.A., Nypaver, M., Poropatich, R., *et al.* (2002). Clinical applications in telemedicine/telehealth. *Telemed J e-Health*, **8**, 13–34.

Kundel, H.L. (1975). Peripheral vision, structured noise and film reader error. *Radiol*, **114**, 269–273.

Kundel, H.L., Nodine, C.F., Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol*, **13**, 175–181.

Kundel, H.L., Nodine, C.F., Krupinski, E.A. (1989). Searching for lung nodules: visual dwell indicates locations of false-positive and false-negative decisions. *Invest Radiol*, **24**, 472–478.

Manning, D.J., Gale, A., Krupinski, E.A. (2005). Perception research in medical imaging. *Br J Radiol*, **78**, 683–685.

Metz, C.E. (2006). Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *JACR*, **3**, 413–422.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

Nodine, C.F., Kundel, H.L. (1987). Using eye movements to study visual search and to improve tumor detection. *RadioGraphics*, **7**, 1241–1250.

Nodine, C.F., Mello-Thoms, C. (2000). The nature of expertise in radiology. In Beutel, J., Van Metter, R., Kundel, H. (eds). *Handbook of Medical Imaging, Vol. 1: Physics and Psychophysics*. Bellingham, WA: SPIE Press, pp. 859–894.

Samei, E., Flynn, M.J., Kearfott, K.J. (1997). Patient dose and detectability of subtle lung nodules in digital chest radiographs. *Health Physics*, **72**(6S).

Smith, M.J. (1967). *Error and Variation in Diagnostic Radiology*. Springfield, IL: Charles C. Thomas.

Tuddenham, W.J. (1962). Visual search, image organization, and reader error in Roentgen diagnosis: studies of psychophysiology of Roentgen image perception. *Radiol*, **78**, 694–704.

Tuddenham, W.J. (1963). Problems of perception in chest roentgenology: facts and fallacies. *Radiol Clin North Am*, **1**, 227–289.

Wagner, R.F., Metz, C.E., Campbell, G. (2007). Assessment of medical imaging systems and computer aids: a tutorial review. *Acad Radiol*, **14**, 723–748.

Weinstein, R.S., Descour, M.R., Liang, C., *et al.* (2001). Telepathology overview: from concept to implementation. *Human Path*, **32**, 1283–1299.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

PART I

HISTORICAL REFLECTIONS AND THEORETICAL FOUNDATIONS

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

2

A short history of image perception in medical radiology

HAROLD KUNDEL AND CALVIN NODINE

> "Offering an account of the past, in disciplinary histories as in ethnic and national ones, is in part a way of justifying a contemporary practice. And once we have a stake in a practice, we shall be tempted to invent a past that supports it."
>
> K. A. Appiah (Appiah, 2008)

2.1 FOREWORD

Medical radiology is a practical field in which images are produced primarily for the purpose of making inferences about the state of health of people. Research in radiology is also practical. Historically, imaging physicists have concentrated on developing new ways to visualize disease and on improving image quality. They have worked on the development of psychophysical models that express mathematically how observers respond to basic properties of displayed images such as sharpness, contrast, and noise. Radiologists have concentrated on image interpretation, which is using images for diagnosis, follow-up, staging, and classification of disease. Image perception, which is the process of acquiring, selecting, and organizing visual information, has generally been neglected perhaps because radiologists take for granted their ability to make sense of the patterns in images. Research in image perception has been motivated by two factors: first, the realization that human factors are a major limitation on the performance of imaging systems and second, the appreciation of the extent of human error and variation in image interpretation. Radiologists certainly are surprised when they discover that they either missed a lesion or saw one that really wasn't there.

This chapter will trace the study of perception and psychophysics as it has unfolded in books and journal articles. We will concentrate on observer error and variation, which has been a major stimulus for the development of a body of statistical methodology known as receiver operating characteristic (ROC) analysis and for attempts at understanding the perceptual basis for image interpretation and reader error. The chapter is based in part on material used by one of us (HK) for a talk given in 2003 at the Medical Image Perception Society (MIPS) in Durham, NC. It reiterates material already published in the *Journal of the American College of Radiology* (Kundel, 2006). Manning, Gale, and Krupinski (Manning, 2005), as well as Eckstein (Eckstein, 2001), also have published histories of medical image perception. A review of research and development in diagnostic imaging by Doi (Doi, 2006) contains observations about image perception and Metz (Metz, 2007) has written a tutorial review of ROC analysis that is considerably more detailed and authoritative than the material presented here.*

2.2 FLUOROSCOPES AND FLUOROSCOPY: A LESSON IN OPTIMIZING IMAGE SYSTEM PERFORMANCE

One of the earliest articles about visual perception in radiology was a discussion of visual physiology and dark adaptation in fluoroscopy by Béclère (Béclère, 1964). Béclère's article was published in 1899 but it was not until 1941 that the impact of dark adaptation on the visibility of details at fluoroscopy was seriously studied. A radiologist, W. Edward Chamberlain, working with the medical physicist George Henny, used the phantom developed by Burger and Van Dijk (Burger, 1936) to measure contrast-detail curves for fluoroscopic screens. They came to the conclusion that although the fluoroscopic screens in use at the time were technically almost equal in sharpness and contrast to images on X-ray films, the decrease in visual acuity and intensity discrimination of the retina at low brightness levels "render the available sharpness and contrast more or less invisible." Chamberlain presented the results in the Carman Lecture at the Radiological Society of North America (RSNA) (Chamberlain, 1942) and suggested that a device called an image intensifier that had been patented recently by Irving Langmuir of the General Electric Research Laboratories could provide a technological solution to the visibility problem. The subsequent development of the image intensifier (Coltman, 1948) vastly improved fluoroscopy and facilitated the development of cineradiography, cardiac catheterization, and interventional radiology.

2.3 THE PERSONAL EQUATION: OBJECTIVELY EVALUATING IMAGE SYSTEM PERFORMANCE

Chamberlain was also involved in the first extensive study of error and variation in radiology. In 1946 the United States

* The history of perception research in radiology as seen through the eyes of two participants is biased by our own experiences and by our tunnel vision. We apologize in advance to those participants in the historical events recorded here whose contribution was slighted, misinterpreted, or not mentioned. Remember there are errors of omission and commission in image interpretation and in recalling history. Feel free to inform us about our error since that is the way that we learn and eventually become experts.

The Handbook of Medical Image Perception and Techniques, ed. Ehsan Samei and Elizabeth Krupinski. Published by Cambridge University Press.
© Cambridge University Press 2010.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

Table 2.1 *Between-observer disagreement. The number of cases read as negative for tuberculosis (neg) by the first reader that were read as positive (pos) by the second reader. The percentage between-observer disagreement is calculated as 100*neg/pos.*

Readers	Neg/pos readings	Percentage inter-observer disagreement
N/M	21/62	34
O/M	19/62	31
P/M	27/62	43
Q/M	11/62	18
Average	19/62	31

Table 2.2 *Within-observer disagreement. The number of cases read as negative for tuberculosis (neg) on a second reading after receiving an initial positive reading (pos). The percentage within-observer disagreement is calculated as 100*neg/pos.*

Readers	Neg/pos readings	Percentage intra-observer disagreement
M	18/118	10
N	4/59	7
O	14/83	17
P	39/96	41
Q	22/106	21
Average	19/92	21

Public Health Service (USPHS) and the Veterans Administration (VA) initiated an investigation of tuberculosis case finding by the newly developed technique of photofluorography. The VA had the responsibility of evaluating induction and discharge chest radiographs on millions of World War II veterans and wanted to use the best of four imaging techniques available for chest screening: 14 by 17 inch celluloid films, 14 by 17 inch paper negatives, 35 mm photofluorograms, and 4 by 10 inch stereophotofluorograms. A "Board of Roentgenology" chaired by Chamberlain and consisting of two radiologists, three chest specialists, and a statistician was convened to address the issue. They designed a study in which five readers independently interpreted four sets of 1,256 cases radiographed using each of the techniques. After a lapse of at least two months, the 14 by 17 inch films were interpreted a second time. The results published in 1947 (Birkelo, 1947) in the *Journal of the American Medical Association* (JAMA) were inconclusive because the variation among readers was greater than the differences among the techniques. The disagreement between pairs of readers averaged about 30% and within pairs of readers about 20%. Tables 2.1 and 2.2 contain brief extracts of the extensive results.

Tables 2.1 and 2.2 illustrate not only the extent of reader disagreement but also the awkward method for summarizing the results. The investigators lacked statistical tools to characterize these data. The JAMA article was accompanied by an editorial (Editorial, 1947) with the title "The personal equation in the interpretation of a chest roentgenogram," which expressed astonishment at the magnitude of observer disagreement and stated: "These discrepancies demand serious consideration." Indeed, the publication of the USPHS–VA study led to a flurry of activity that is succinctly described by two of the major participants, the radiologist L. Henry Garland (Garland, 1949) and the project statistician Jacob Yerushalmy (Yerushalmy, 1969).

The phrase "the personal equation" goes back to 1796 when the British Astronomer Royal, Nevil Maskelyne, found that his observations of the time that a certain star crossed the meridian were different from those of his assistant (Stigler, 1968). The transit time was used to set ship navigational clocks and although the error of eight-tenths of a second only translated into one-quarter of a mile at the equator, it was important to an astronomer. Maskelyne and his assistant tried to get their measurements to agree but after repeated attempts they failed. He fired the assistant! Twenty years later, while writing his *Fundamenta Astronomiae*, Friedrich Bessel found Maskelyne's account and did some experiments that also showed observational variation among astronomers. He tried, unsuccessfully, to develop "personal equations" to adjust for the differences between observers.

$$\text{John's value} = \text{Jane's value} + \text{bias correction} \tag{2.1}$$

It is ironic that in 1994 an editorial (Editorial, 1994) in the *New England Journal of Medicine* (NEJM) accompanying an article with the title "Variability in radiologists' interpretation of mammograms" (Elmore, 1994) expressed similar sentiments to those in the JAMA editorial. It is distressing that the authors either ignored or were unaware of 50 years of research on observer variation in radiology.

The results of the VA chest screening study eventually were expressed as under-reading and over-reading. A number of follow-up studies were designed "with the hope of discovering the components responsible for this variability" (Yerushalmy, 1969). Two groups of the radiologists that participated in the studies were designated CRN for Chamberlain, Rigler, and Newell and GMZ for Garland, Miller, and Zwerling. The CRN results (Newell, 1954) were published in a paper titled "Descriptive classification of pulmonary shadows: a revelation of unreliability in the roentgen diagnosis of tuberculosis." The GMZ results (Garland, 1949) were summarized by L. Henry Garland in his presidential address to the RSNA in 1948 titled "On the scientific evaluation of diagnostic procedures." His 1959 update of error in radiology and medicine in general is often quoted to support a statement that radiologists disagree with each other 30% of the time (Garland, 1959). Diagnostic unreliability has not gone away. Similar observations about disagreement (Felson, 1973; Gitlin, 2004; Goddard, 2001) are made whenever it is specifically studied.

Garland could not explain the observed variability. He classified the errors using a taxonomic approach that was later elaborated by Smith (Smith, 1967) and updated by Renfrew *et al.* (Renfrew, 1992; see Table 2.3 in Section 2.5). The GMZ group also studied reading strategies, which included dual reading and trying to control the attitude of the reader. The use of dual reading as an error-compensating mechanism was the major practical suggestion that resulted from the USPHS–VA study (Yerushalmy, 1950). Garland wrote the following about the

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

attitude studies: "In studying the problem, the group was very conscious of the penalty in the form of over-reading which must be paid for the advantage of a reduction in under-reading." They had hit upon attitude or bias toward a particular outcome as a source of variability and recognized that it influenced the ebb and flow of true and false positives but they could not deal with it because they lacked an adequate model. That model was found in signal detection theory and a radiologist, Lee Lusted (Lusted, 1968), was largely responsible for its introduction into both radiology and the entire medical community.

2.4 RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS

2.4.1 The introduction of signal detection theory into radiology

The theory of signal detectability was developed by mathematicians and engineers at the University of Michigan, Harvard University, and the Massachusetts Institute of Technology partially as a tool for describing the performance of radar operators. Lee Lusted was exposed to the concepts of signal and noise in 1944 and 1945 while working in the radio research laboratory at Harvard University (Lusted, 1984). In 1954, as a radiology resident at the University of California in San Francisco (UCSF), he was introduced to the problem of observer error when he participated in one of the film reading studies supervised by Yerushalmy and Garland. Apparently his mind was prepared for a logical leap when in 1956 and 1957 he encountered a plot of percentage false negatives against percentage false positives in the laboratory of W. J. Horvath, who was responsible for optimizing the performance of the Cytoanalyzer, which was a device for automatically analyzing Papanicolaou smears (Horvath, 1956). At that time Lusted plotted a "performance curve" for chest X-ray interpretation. In 1959 he showed the curve reproduced in Figure 2.1 in the Memorial Lecture at the RSNA and published it in 1960 (Lusted, 1960). This was the first published example of an ROC curve for performance data from radiology.

Although Figure 2.1 shows a plot of false negatives against false positives, the usual convention, shown in Figure 2.2, is to plot true positives against false positives.

Lusted (Lusted, 1969) saw the ROC curve as a useful tool to accomplish two things: first, to use a parameter such as the area under the curve (AUC) as a single figure-of-merit for an imaging system and second, to decrease the observed variability in reports about images by separating the intrinsic detectability of the signal, which is a sensory variable, from the decision criteria, which is a matter of judgment. He stated: (Lusted, 1969) "It is very difficult for a human observer to maintain a constant decision attitude over a long period of time. This is a possible explanation for the finding that a radiologist will disagree with his own film interpretation about one out of five times on a second reading of the same films." He wrote a very influential book, *Introduction to Medical Decision Making* (Lusted, 1968), and went on to become a founding member of the Society for Medical Decision Making and the first editor of the journal, *Medical Decision Making*.

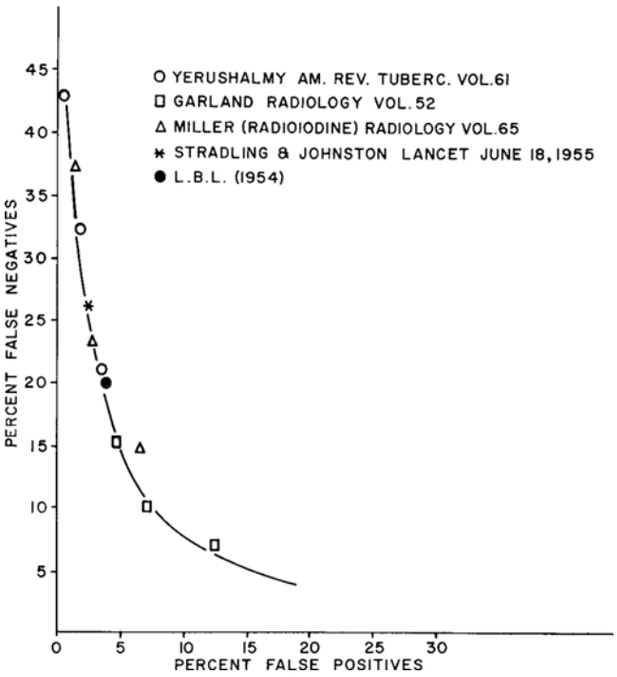

Figure 2.1 The "operating characteristic" curve showing the reciprocal relationship between percentage false negatives and percentage false positives. Most of the data were from studies of the interpretation of photofluorograms for tuberculosis. From Lusted (Lusted, 1960) with permission.

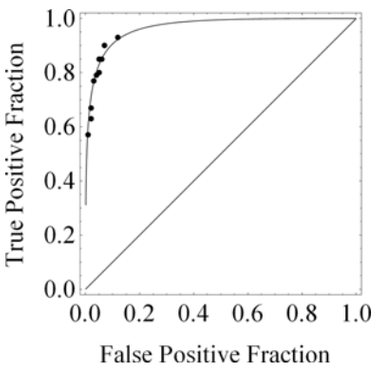

Figure 2.2 A conventional receiver operating characteristic curve showing the reciprocal relationship between the fraction of true positives and the fraction of false positives. The data points are the same as those in Figure 2.1. The curve is a binormal curve with an area under the curve (AUC of Az) of 0.87 that was fit by inspection to the data points.

Signal detection theory is a psychophysical model that describes an observer's response in terms of some known or estimated distribution of a signal and noise in the stimulus. The theoretical foundations of signal detection theory laid out in a book by Green and Swets was originally published in 1966 (Green, 1966) and reprinted with revisions in 1974 (Green, 1974). ROC analysis, which is derived from the theory of signal detectability, has become a powerful tool in visual systems evaluation. It turns out that some of the variability among observers can be reduced by applying a signal detection theory model. Perhaps

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-51392-0 - The Handbook of Medical Image Perception and Techniques
Edited by Ehsan Samei and Elizabeth Krupinski
Excerpt
More information

the personal equation should be written in terms of the linear equation that describes an ROC curve in normal deviate space.

ROC index of detectability
$$= z\,(\text{true positive fraction}) - z\,(\text{false positive fraction}) \qquad (2.2)$$

where z is the normal deviate.

2.4.2 Early studies of ROC analysis in radiology

ROC analysis was not embraced immediately by the image technology evaluation community. The method was unfamiliar and practical examples illustrating experimental design, data collection using rating scales, and ROC parameter calculation were not readily available. Some early studies used an ROC parameter, the index of detectability, d', read from tables published in a book about signal detection theory (Elliott, 1964), to obtain a single estimate of performance from true positive and false positive pairs (Kuhl, 1972; Kundel, 1968). The data lacked estimates of variance and the use of d' in the absence of information about the complete ROC curve made assumptions about the ROC parameters that may not have been justified (Metz, 1973a).

The situation was improved when Dorfman and Alf (Dorfman, 1969) at the University of Iowa published a method using maximum likelihood for estimating the parameters of the ROC curve. Much of the subsequent development of statistical methodology was based on this work. In the 1970s David Goodenough, Kurt Rossmann, and Charles Metz (Goodenough, 1972, 1974; Metz, 1973b) at the University of Chicago demonstrated the use of ROC analysis in the evaluation of film-screen combinations for standard radiography. A number of articles describing the value and the use of ROC technique were published in the 1970s (Lusted, 1978; McNeil, 1975; Metz, 1978).

In 1979 Swets *et al.* (Swets, 1979) reported the results of a multi-institutional study, supported by the National Cancer Institute (NCI), comparing the accuracy of radionuclide scanning and computed tomography (CT) for detecting and classifying brain tumors. The study was more important as a demonstration of the potential power of the ROC methodology for technology evaluation in a clinical environment than as a comparison of two imaging methods. The methodology for evaluating "diagnostic systems" using ROC analysis was described in a book by John Swets and Ronald Pickett (Swets, 1982) that laid the groundwork for future developments in statistical methodology. A FORTRAN version of the Dorfman–Alf computer program was published in the book as an appendix. This became a prototype for the subsequent development in the 1980s by the group at the University of Chicago of a very influential ROC analysis software package called ROCFIT. It was superseded in the 1990s by a new package called ROCKIT.

2.4.3 Developing methods for the statistical analysis of ROC data

A test of whether the difference between two values of the area under the ROC curve is due to a real difference in the imaging techniques that yielded the values or just due to chance can be done by calculating a critical ratio (CR), denoted z, which is the ratio of the observed difference ($AUC_1 - AUC_2$) to the standard error ($SE_{(diff)}$) of the difference. The CR is then used to estimate the probability that the difference is real (or statistically significant) (Hanley, 1983).

$$z = \frac{AUC_1 - AUC_2}{SE_{(diff)}} \qquad (2.3)$$

The AUC can be calculated using the procedure of Dorfman and Alf (Dorfman, 1969). Calculating the $SE_{(diff)}$ is not as straightforward. Swets and Pickett (Swets, 1982) proposed a model that included estimates of the variability due to case sampling, reader sampling (between reader variability), reader inconsistency (within reader variability), and the multiple correlations between cases and readers. They also presented a methodology for approximating the estimates. In 1992 Dorfman and Berbaum from the University of Iowa and Metz from the University of Chicago published a method for analyzing rating scale data using a combination of the Dorfman–Alf method for calculating ROC parameters and the classical analysis of variance (ANOVA) (Dorfman, 1992). The so called multi-reader multi-case (MRMC) or Dorfman, Berbaum, and Metz (DBM) method separates case variance from within and between reader variance, providing a method for deciding whether any observed differences are due to the readers or to the cases.

Recent work on methodology has focused on improving techniques to account for variance (Beiden, 2001) and on accurate estimates of statistical power (Chakraborty, 2004; Hillis, 2004, 2005; Obuchowski, 2000a).

2.4.4 ROC analysis becomes a standard method for technology evaluation

In 1989 four articles that reviewed the state of the art of ROC analysis were published by the groups that were most active in methodological development: Berbaum *et al.* from the University of Iowa (Berbaum, 1989), Gur *et al.* from the University of Pittsburgh (Gur, 1989), Hanley from McGill University (Hanley, 1989), and Metz of the University of Chicago (Metz, 1989). The fact that four reviews were published indicates the growing interest in ROC analysis as a methodology for imaging technology evaluation.

A count of the articles in six radiology journals indexed by PubMed that use the phrase "ROC" in either the title, the abstract, or the keywords is shown in Figure 2.3.

There is a steady increase in the number of citations since 1974. Note that the jump in 1988 may be due to an increase in publications but may also be an artifact caused by the addition to the database of abstracts and the keyword "ROC".

There has been increased use of ROC analysis for technology evaluation and a steady development of the methodology. There is now a new generation of review articles (Metz, 2007; Obuchowski, 2005) and ROC analysis is even beginning to appear in statistics textbooks (Zhou, 2002).

© in this web service Cambridge University Press www.cambridge.org