Finding order in complexity: Themes from the career of Dr. Robert F. Wagner

Kyle J. Myers*

Division of Imaging and Applied Mathematics, OSEL, CDRH, FDA
WO62-3118, 10993 New Hampshire Avenue, Silver Spring, MD 20993-0002

ABSTRACT

Over the course of his long and productive career, Dr. Robert F. Wagner built a framework for the evaluation of imaging systems based on a task-based, decision theoretic approach. His most recent contributions involved the consideration of the random effects associated with multiple readers of medical images and the logical extension of this work to the problem of the evaluation of multiple competing classifiers in statistical pattern recognition. This contemporary work expanded on familiar themes from Bob's many SPIE presentations in earlier years. It was driven by the need for practical solutions to current problems facing FDA'S Center for Devices and Radiological Health and the medical imaging community regarding the assessment of new computer-aided diagnosis tools and Bob's unique ability to unify concepts across a range of disciplines as he gave order to increasingly complex problems in our field.

Keywords: R.F. Wagner, image evaluation, objective assessment, decision theory, CAD

1. BACKGROUND

Robert F. Wagner began his career at the FDA in the early 1970s. He arrived with a recent Ph.D. in nuclear physics and a charge to build a medical imaging program at a time of tremendous innovation in this early field. Bob approached this charge in a fashion that was to be a standard modus operandi for his professional life: he sought out experts in related fields (Otto Schade and other TV picture quality gurus of the time, in this case[1]) and worked to translate and adapt relevant concepts to his area of application. His first SPIE paper[2] laid the groundwork for what was to come, as he summarized the assortment of image quality indexes for radiographic film-screen combinations in existence at the time and wondered publicly – could they be resolved? For the next 35 years, he endeavored to build the science of image assessment, tackling first the issue of quantum noise and film mottle in medical images and taking on problems of greater complexity and clinical relevance over time. From the outset, he wrote and spoke prolifically, allowing great transparency into his forming ideas and methods. His great passion for technical exchange played a significant role in the formation of the SPIE Medical Imaging Conference. Each year he came with his latest update, giving over sixty SPIE Medical Imaging conference papers in all. It was this meeting that singularly afforded Bob the opportunity to nurture life-long friendships with colleagues, discuss his latest insights, and work toward the development of consensus.

2. THE ASSESSMENT OF QUANTUM-LIMITED MEDICAL IMAGES

As described in the companion paper in this proceedings volume by David Brown,[3] Bob's earliest work considered the quantum-limited performance of medical imaging systems.[4,5] In their 1985 Physics in Medicine and Biology paper,[6] Wagner and Brown presented a "Unified SNR analysis of medical imaging systems," which laid out a general information-theoretic form for imaging performance for all the major medical imaging modalities of the time. This landmark paper was one of the first of many significant works in Bob's career in which he revealed an order across complex and distinct phenomena, in this case, the commonalities found in the assessment of the detectability of simple structures in quantum-limited images across the wide variety of physics processes that can be harnessed in the formation of medical images.

*kyle.myers@fda.hhs.gov, (301) 796-2533.

Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment,
edited by Berkman Sahiner, David J. Manning, Proc. of SPIE Vol. 7263, 72630H
© 2009 SPIE · CCC code: 1605-7422/09/$18 · doi: 10.1117/12.817794

Bob's unified SNR approach to medical imaging assessment was in terms of the best possible observer, the Bayesian observer, for simple detection of classification tasks. The question naturally arises regarding how well human observers are able to realize these upper-bound performance predictions. Art Burgess, then of the University of British Columbia, Canada, spent some sabbatical months with Bob in 1980, during which time they submitted a paper to Science, along with coauthors Bob Jennings and Horace Barlow, on the detection efficiency of human observers relative to the ideal observer.[7] This first comparison of human and ideal observers in quantum-limited images showed that humans were remarkably efficient, operating at about 50% efficiency for simple detection tasks in white noise.

Bob continued to work on problems related to machine classifiers and human observer performance for the rest of his career. He was tremendously interested in understanding for what tasks humans were efficient and when they were not, how ideal observer models might be "handicapped" to better approximate human performance (and how such handicaps might be motivated by known mechanisms of the human visual system), and when computer-assist devices might improve human performance. I first got to know Bob through discussions of these topics during my graduate school days at the University of Arizona in the early 1980s, when Harry Barrett and I were investigating the impact of high-frequency noise on human detection performance. We were greatly influenced by Bob's work on the use of ideal-observer SNRs for image quality, and Art Burgess' results on human signal detection efficiency in CT noise.[8] We developed an experiment that allowed the probing of the impact of noise correlations on human detection performance for higher noise powers, and found that the human efficiency dropped precipitously as the power of the noise increased. Furthermore, we found that the data were well modeled by a non-prewhitening ideal observer, that is, one suffering from a reconstruction penalty, as it was referred to in Wagner and Brown's paper. Our line of experimental inquiry, which led to the conception of a channelized ideal observer[9] as an alternative to the non-prewhitening matched filter, was greatly inspired by the work of Bob and his collaborators.

In 1987, not long after I graduated from the University of Arizona, I had the opportunity to come to the Center for Devices and Radiological Health to work as Bob's junior research colleague. It was an opportunity I couldn't pass up! Since that time, I have had many people relate similar stories of how generous Bob was with the time he spent getting to know them and their work during their student days. He was extraordinarily generous in terms of the time he spent with students at their conference posters, taking an interest in their work and offering encouragement early in their careers. Bob had a tremendous heart for young people. Back at the office he preferred to supervise one young investigator at a time, either a grad student or fairly recent graduate, who would have the amazing opportunity to work one-on-one with Bob for hours daily over the course of several years. I consider myself truly blessed for having been a link in that chain of Bob's young investigators.

3. THE NEXT LEVEL OF COMPLEXITY: BIOLOGICAL VARIABILITY

Realistic images have complicated background structures, and may have random signals, too. Thus, once Bob had unified image evaluation for the "signal-known-exactly, background-known-exactly" (SKE/BKE) paradigm, he and his colleagues turned to more complicated and realistic imaging tasks involving random backgrounds and variable signals. Much of the motivation for the early work on image performance evaluation in the presence of a lumpy background (see the companion paper in this proceedings by H.H. Barrett[10]) came from passionate discourse between Bob, Harry, and others regarding the sometimes surprising and counterintuitive results of system optimization studies for tasks involving known signals and backgrounds. Bob and Harry were coauthors on several papers related to this topic, extending the decision-theoretic framework of image assessment to allow for statistical backgrounds, and in the process, demonstrating Bob's remarkable capacity for engaging his colleagues in passionate and sometimes heated debate, while maintaining mutual respect and collegiality.

Bob took great pride in those times in his career when he challenged someone's viewpoint, even if in the end it was revealed that, for example, because a study he had published was somewhat unrepresentative of clinical practice it had resulted in peculiar conclusions. In this case, the SKE/BKE paradigm was found to be flawed, but the pay off in terms of new models for tasks and evaluation strategies was enormous. A number of unifying outcomes from this period in the field of medical image assessment continue to be important today, largely as a result of Bob's influence through his own research and his impact on the work of others. In particular, it is now widely appreciated that the evaluation of a medical imaging system must consider tasks involving statistical backgrounds. While this reality brings complexity to the image evaluation process, a beautiful order can be found in the resulting extensions to the expressions for the SNR of the ideal observer and the ideal Hotelling observer in the presence of background variability.[11]

4. ARTIFACT-LIMITED IMAGES

Bob Wagner first considered the role of image reconstruction in tomographic imaging in the late 1970s, refuting the claims of some investigators who believed they could greatly reduce the dose of CT through efficient algorithms.[3,12] Bob and his colleagues demonstrated that CT was inherently a high-dose modality (relative to the planar x-ray imaging methods that preceded it). Even so, the potential for effective limited-angle tomography appealed to Bob, who championed methods for the task-based assessment of potential geometries in CT and MRI.

As a demonstration effort, Bob collaborated with Ken Hanson of Los Alamos National Laboratory on a series of papers in the early 1990s on the evaluation of image reconstruction algorithms. (See companion proceedings in this volume by Ken Hanson.[13]) A variety of planar CT image acquisition geometries, noise levels, image reconstruction algorithms, and tasks were considered, with both human and algorithmic observers employed. This work demonstrated the applicability of the same signal-detection theoretic approach to the evaluation of artifact-limited images that Bob first championed for simple quantum-limited images. Furthermore, the evaluation of artifact-limited images was shown to require tasks with randomness in the signals and/or backgrounds (biological variability), because artifacts are only problematic for the ideal observer when they are randomized – the limited-angle tomographic image of a nonrandom object may be "artifacty" but it is still known exactly.

4. THE COMPLICATION OF READER VARIABILITY

The introduction of full-field digital mammography (FFDM) as an alternative to film-screen mammography (FSM) was greatly anticipated in the 1990s by the medical imaging community, following years of investment by industry and the NIH's National Cancer Institute. However, it became clear that a simple side-by-side comparison of FSM and FFDM images would not be a feasible means of demonstrating the effectiveness of this new FFDM technology, as there was no consensus on the clinical meaning of the resulting differences in image values. The standard physical performance measures, MTF and NPS, and the associated DQE, which had come to be relied upon in the evaluation of FSM, also raised questions for some regarding the clinical significance of the differences in these measures found between these alternative technologies. Hence, the medical community pushed for clinically relevant reader study. Complicating this undertaking was the quantitative measurement of reader variability in mammography that was published by Beam *et al.* in 1996.[7]

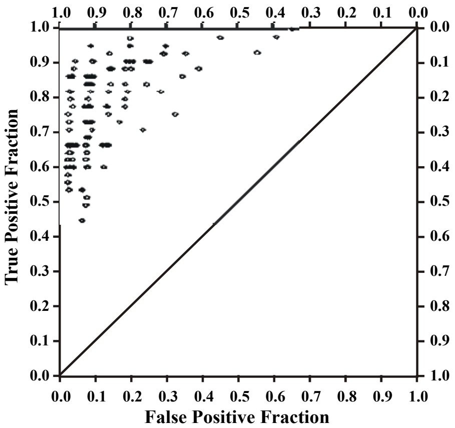

Figure 1. Sensitivity-specificity pairs for 108 US mammographers in the study by Beam *et al.*[14] (Adapted with permission from Archives in Internal Medicine, Vol. 156, p 209-213; Copyright 1996 American Medical Association. All rights reserved.)

The complication of reader variability is dramatically demonstrated in Figure 1, adapted by Bob from the Beam *et al.* publication. The figure presents the results of a study of 108 US mammographers interpreting a common set of mammograms. This plot of the sensitivity-specificity pairs of each radiologist was a favorite figure of Bob's. He used it frequently to make the point that there is a range in both reader skill and reader aggressiveness, or mindset, and thus there is no single operating point for the mammographers participating in this study. A reader study to compare the performance of two imaging modalities like FSM *vs.* FFDM must contend with the implication of this additional source of variability, on top of the variability that would be expected from the measurement noise and biological (or case) variability.

Figure 2 illustrates the multiple factors we have discussed that may contribute to the spread in an observer's decision variable in an objective imaging performance analysis. A figure like this appeared in ICRU Report 54, Medical Imaging – The Assessment of Image Quality,[15] of which Bob was a principal coauthor. This document was commissioned by the International Commission on Radiation Units and Measurements to present the best understanding of the current framework, based on statistical decision theory, by which imaging system performance could be measured, optimized, and compared. The report advocated the use of receiver operating characteristic (ROC) curve analysis for measurement of task performance by observers, but did not address the estimation of uncertainty in the resulting performance measures. At the time of the report's publication, no software tools were available for analyzing these contributions to the uncertainty in a figure of merit.

To account for the kinds of variability seen in the study by Beam *et al.,* a number of statistical solutions[16-20] and related software[21,22] were released in the 1990s, all involving so-called multiple-reader multiple case (MRMC) ROC analysis to account for the multiple components of the variance (cases, readers, and their interactions) observed in ROC studies. Bob recruited a bright young Russian mathematician, Sergey Beiden, to work with him on the development of a non-parametric statistical analysis method based on bootstrap resampling. Over the course of the next several years, Bob and Sergey released a series of papers addressing various MRMC issues related to experimental design and a problem's variance structure,[23-25] including the impact of continuous *vs.* categorical rating data,[26] the difference in reader variance without *vs.* with CAD,[24] and an analysis of the components of variance when CAD is used in an independent *vs.* sequential reading paradigm.[27] The reference list shows that he carried out this work with a large number of prestigious collaborators. He was particularly proud of the joint paper on the inter-comparison of MRMC analysis methods coauthored by Nancy Obuchowski and other leaders in the field,[28] as well as the tutorials (Bob liked to call them "pre-guidance") he published with Greg Campbell (head of CDRH's Division of Biostatistics) and Charles Metz.[29-31] Charles was a singular source of ideas, wisdom, and feedback on these topics. Bob might close his door, but his big laugh still spilled out under the door and through the walls during their frequent late afternoon conversations.

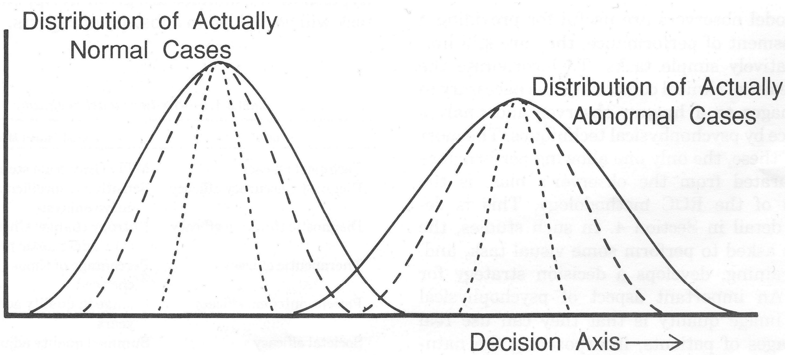

Figure 2. The spread of an observer's decision variable owing to the following factors: random noise fluctuations (- - -), noise and biological variability (– – –), and that from noise, biological variability, artifacts and observer variability (——).

Bob gave numerous presentations inside FDA to develop an awareness and appreciation for the MRMC methodology and its importance for imaging device comparison studies as he pushed for the adoption of this paradigm by the agency. He argued that the fully-crossed MRMC experimental design, where every reader interprets every case from two modalities under comparison, has the most statistical power for a given number of readers and a given number of cases with verified truth; thus, it's least demanding of these resources, or "least burdensome," in the language of the agency. Most notably, in March 2001 he gave a tutorial on multivariate ROC analysis to the FDA/CDRH advisory panel meeting convened to consider the data submitted to support the first premarket application for an FFDM imaging system,[32] followed by an update when an advisory panel convened to consider the data submitted to support the first premarket application for a lung nodule computer-aided detection (CAD) device.[33] He ardently sought presentation slots at these very public meetings, knowing they offered tremendous opportunities for education of a wide audience, including the advisory panel and all the members of the audience who would be making similar petitions to the agency in the future.

5. COMPUTER-ASSIST DEVICES AND STATISTICAL LEARNING MACHINES

Early on, Bob recognized the analogy that could be drawn between the sources of variability in an MRMC study involving human readers and the sources of variability in a study evaluating the performance of a computer-aided detection algorithm. For a given type of machine classifier (fixed architecture), different training sets will result in different particular algorithms (different weights, for example), resulting in different levels of performance. For both kinds of observers, man and machine, differences in observer skill can be attributed in part to differences in the range of cases in the training set. Of course, CAD algorithms can have different mindsets, or thresholds, too, just as readers can. And finally, the range of test cases impacts mean performance and its variability similarly for both kinds of observers as well. Thus Bob's recent presentations routinely made the point that resampling is essential for the assessment of statistical learning machines, because there is uncertainty from the finite test set, the finite training set, and their interactions.

Figure 3 shows two "learning curves" from a 1994 SPIE Medical Imaging paper Bob wrote "On combining a few diagnostic tests or features."[34] The x axis plots the number of patients used to train a simple Bayesian classifier (linear in this uncorrelated Gaussian statistics problem). The y axis plots the mean estimate of the area under the ROC curve over a set of 250 Monte Carlo trials; the testing case set is very large. With this simple graph Bob was able to make several key points, including: limited training data negatively impacts the ability of the classifier to reach its performance potential; and the addition of an informative feature can improve classifier performance. Many additional plots can be found in this paper, illustrating what happens in higher dimensions, or when noisier or more correlated features are added, for example.

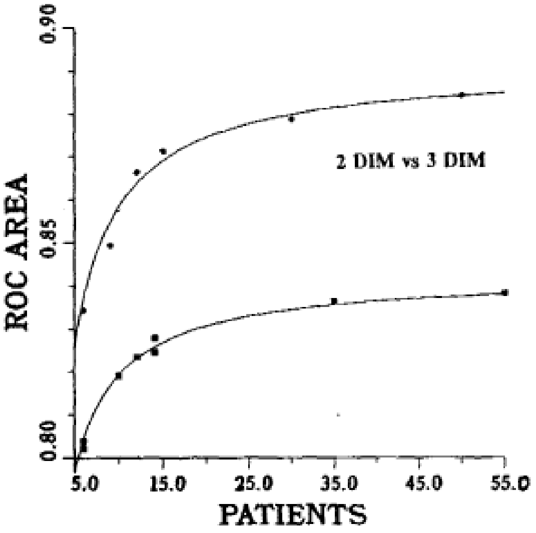

Figure 3. Mean performance of a linear classifier with 2 and 3 uncorrelated, Gaussian features as a function of the number of training cases.

When data like those shown in Fig. 3 are replotted against the reciprocal of the number of training samples, we obtain plots like those shown in Figure 4.[35] These plots are examples of what Bob referred to as "antler diagrams," thanks to the inclusion of a second set of performance data with an inverse bias to that illustrated in Fig. 3. Specifically, Fig. 4 demonstrates the positive bias in the performance estimate one finds when a classifier is trained and tested on the same samples – resubstitution, the upper antlers – as opposed to the negative bias that happens when training and testing on independent samples – the lower antler.

Unlike Figure 3, Figure 4 also includes error bars indicating the uncertainty in the performance estimates for each condition. Not long after these data were presented to the SPIE community in 1997, Bob began calling for performance estimates of CAD algorithms that accounted for the training uncertainty in the estimated performance of a CAD device or machine algorithm. While many algorithm designers would argue that they were only interested in the performance of their particular device, Bob was unrelenting in his campaign on this issue. He strongly believed that a full evaluation of the *technology,* as he referred to it, required an analysis of the uncertainty owing to both the finite number of training and testing samples. The particular training set is a random effect that needs to be accounted for. In fact, Bob found that the finite-sample training variance frequently dominates the uncertainty in the performance difference between competing classifiers.[36]

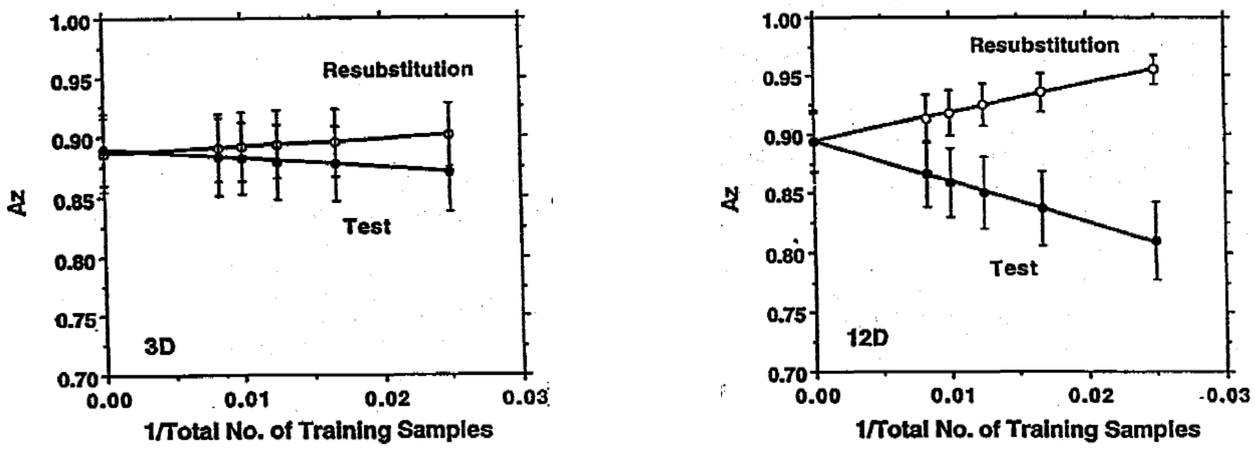

Figure 4. Mean area under the ROC curve, Az, and estimates of its uncertainty for a linear classifier in a Gaussian-distributed multiple-feature problem, as a function of the reciprocal of the number of training samples. The upper plot in each graph shows the performance bias when the classifier is tested on the same cases used in its training. The lower plot shows performance for a classifier trained and tested on independent cases. The left plot is for a 3-D feature space; the right plot shows the results for a 12-D feature space. Bob referred to plots like these as antler diagrams.

Bob's earliest work on classifier performance estimation focused on fairly low-dimensional problems inspired by the algorithms being developed for applications in mammography and other radiological data sets. There might be a million pixels in a mammogram, but there is a great deal we know regarding the spatial correlations within and across organs, and the image features associated with the pathologies to be detected. Early CAD development involved gathering such information from expert radiologists and clinicians and attempting to write code that extracted such features quantitatively. In this way, a CAD algorithm would reduce those million pixels in a mammogram down to a much smaller feature space. The low dimensionality of the reduced features space meant there would be reasonable hope of building a data base of training and testing cases with sufficient numbers of samples for device development and performance estimation.

Most recently, Bob's work had been devoted to the increased complexity found in the assessment of computer-aided diagnosis methods applied to modern multiple biomarker problems: finding genes, classifying proteins, and other so-called "-omics" tasks using microarrays.[37,38] Bob steeped himself in the contemporary literature on statistical learning theory, support vector machines, and the biological discoveries paving the way to the touted future of personalized medicine. As he educated himself, he brought his expertise on image assessment methodologies to new audiences, including the multi-institutional MicroArray Quality Control Consortium,[39] and supported the development of regulatory decision-making paradigms for related products in FDA/CDRH.

The tremendous dimensionality of the data sets in modern multiple biomarker problems raises unique challenges because in this field, unlike medical imaging, information on the relationships between data elements is still being discovered, and therefore the problem of feature selection, or data mining, is significant. Figures 3 and 4 show the performance characteristics of simple linear classifiers; in addition, Bob and his colleagues had investigated more complex classifiers, such as quadratic discriminants and artificial neural nets.[40] Bob knew from these simulations and theoretical investigations that the slopes of the learning curves and the uncertainties associated with the finite training set increase with the dimensionality of the problem and the complexity of the classifier.

Formal definitions of the complexity of a classifier exist, related to the ability of the classifier to partition data sets of given statistical properties.[41] Related to classifier complexity is 1) the amount of data required to train the classifier in order for its performance to approach its large-training-sample performance and 2) the stability of the classifier's performance estimate with respect to different training sets of fixed size. The analogies Bob had drawn earlier regarding readers and training cases, patients and testing cases, and imaging systems and algorithms apply all the more so to the multiple biomarker problem. Bob's research goal was to develop practical assessment tools and paradigms for users of these emerging technologies. He continued to advocate for an estimation of the total uncertainty from both the finite training dataset and the finite testing dataset.[42,43] He laid out a three-stage process of classifier development and assessment, to be followed once preclinical biological, chemical, pharmacogenomic, etc., research had suggested the possible utility of an array of biomarkers:

Level 0: Data-mining
Select features (biomarkers)
Select classifier architecture

Level 1: Pilot Study
Collect new samples
Estimate the mean performance and total uncertainty (from finite trainers and testers)
Estimate how to scale up to a pivotal study that would yield desired confidence levels

Level 2: Pivotal Study
Collect more samples (can incorporate samples from Level 1)
Divide samples into independent training and testing sets for final performance evaluation

Bob hoped that, by following this 3-level approach, fewer blind alleys would stall progress in the microarray biomarker world. In the years to come, as new biomarkers are discovered and related algorithms are investigated to harness their power to personalize medicine, we will be able to witness first hand the impact of Bob's pioneering work on this promising field.

6. FINAL REMARKS

The title of this talk is a slight modification of the title of one of Bob's favorite books, Order out of Chaos, by Ilya Prigogine. When I arrived at the FDA to work with Bob, chaos theory was a hot topic in the physical sciences, and Bob read countless books on the subject. He even arranged to give a multi-installment course on chaos to all the other scientists in the CDRH labs! Such was his enthusiasm for understanding this field and passing on his excitement about it to others. As a new scientist working for Bob, I was amazed by the energy he put into this side project and the way the whole lab stopped work for many hours to enjoy the gift of his time and teaching talent.

Bob relished being able to make sense of complex processes. The body of his professional work is filled with examples where he brought people from separate disciplines together, built bridges between ideas and problem areas, and worked to create a unified framework for understanding the world around him. Bob once wrote:

> *"One of the great discoveries of the post-enlightenment world (~ last 250 years or so) is that the world *out there* is quite complicated. In fact, in my humble opinion it is a mark of an educated person that they have been exposed to the discoveries and understanding of just how complicated the world really is--whether it be in terms of psychology or sociology or economics or medicine or the physical sciences--or anything that is really at all interesting to think about. The great challenge and fun is how people have been able to break complicated things down into bite-sized pieces…"*

This quote truly speaks to Bob's own life – Bob certainly challenged himself, and had fun in the process, as he broke down complicated problems into relatable pieces, and shared his insights and research progress at the SPIE Medical Imaging Symposium each year. He had a unique ability to unify concepts across a range of disciplines, and in doing so, he gave order to increasingly complex problems in our field. Along the way, he challenged us all and gave us a whole lot of fun, too. This meeting, and the field, will greatly miss him.

7. REFERENCES

[1] Wagner, R.F., “Decision theory and the detail signal-to-noise ratio of Otto Schade,” Photog. Sci. Eng. 22(1):41-46 (1978).

[2] Wagner, R.F. and Weaver, K.E., “An assortment of image quality indexes for radiographic film-screen combinations--can they be resolved?” Proc. SPIE 35:83-94 (1972).

[3] Brown, D.G., “Bob’s first decade: in the beginning,” Proc. SPIE 7263-12 (2009).

[4] Wagner, R.F., Weaver, K.E., Denny, E.W., Bostrom, R.G., “Toward a unified view of radiological imaging systems. Part I: Noiseless images,” Med. Phys. 1:11-24 (1974).

[5] Wagner, R.F., “Toward a unified view of radiological imaging systems. Part II: Noisy images,” Med. Phys. 4:279-296 (1977).

[6] Wagner, R.F., and Brown, D.G., “Unified SNR analysis of medical imaging systems,” Phys. Med. Biol. 30:489-518 (1985).

[7] Burgess, A.E., Wagner, R.F., Jennings, R.J., Barlow, H.B., “Efficiency of human visual discrimination,” Science 214:93-94 (1981).

[8] Burgess, A.E., “Statistical efficiency of perceptual decisions,” Proc. SPIE 454:18-26 (1984).

[9] Myers, K.J., and Barrett, H.H., “Addition of a channel mechanism to the ideal-observer model,” J. Opt. Soc. Am. A 4:2447–2457 (1987).

[10] Barrett, H.H., “NEQ: its progenitors and progeny,” Proc. SPIE 7263-14 (2009).

[11] Barrett, H.H. and Myers, K.J., Foundations of Image Science, John Wiley & Sons, Inc., New York (2004).

[12] Wagner, R.F., Brown, D.G. and Pastel, M.S., “The application of information theory to the assessment of computed tomography,” Med. Phys. 6:83–94 (1979).

[13] Hanson, K.M., “Performance-based assessment of reconstructed images,” Proc. SPIE 7263-15 (2009).

[14] Beam, C.A., Layde, P.M., and Sullivan, D.C., “Variability in the interpretation of screening mammograms by US radiologists,” Arch. Intern. Med. 156:209-213 (1996).

[15] International Commission on Radiation Units and Measurements. *Medical Imaging: The Assessment of Image Quality.* ICRU Report 54 (Bethesda, MD: ICRU) (1996).

[16] Dorfman, D.D., Berbaum, K.S., Metz, C.E., “Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method,” Invest. Radiol. 27:723–731 (1992).

[17] Obuchowski, N.A., “Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations,” Acad. Radiol. 2(Suppl 1):S22–S29 (1995).

[18] Toledano, A., Gatsonis, C.A., “Regression analysis of correlated receiver operating characteristic data,” Acad. Radiol. 2(Suppl 1):S30–S36 (1995).

[19] Toledano, A., Gatsonis. C.A., “Ordinal regression methodology for ROC curves derived from correlated data,” Stat. Med. 15:1807–1826 (1996).

[20] Ishwaran, H., Gatsonis, C.A., “A general class of hierarchical ordinal regression models with applications to correlated ROC analysis,” Can. J. Stat. 28:731–750 (2000).

[21] University of Chicago LABMRMC software. Available online at: http://www-radiology.uchicago.edu/krl/roc_soft6.htm Accessed January 31, 2009.

[22] University of Iowa MRMC software. Available online at: http://perception.radiology.uiowa.edu/Software/tabid/109/Default.aspx Accessed January 31, 2009.

[23] Beiden, S.V., Wagner, R.F., Campbell, G., “Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects receiver operating characteristic analysis,” Acad. Radiol. 7:341–349 (2000).

[24] Beiden, S.V., Wagner, R.F., Campbell, G., Metz, C.E., Jiang, Y., “Components-of-variance models for random-effects ROC analysis: The case of unequal variance structures across modalities,” Acad. Radiol. 8:605–615 (2001).

[25] Beiden S.V., Wagner, R.F., Campbell, G., Chan, H-P., “Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis,” Acad. Radiol. 8:616–622 (2001).

[26] Wagner, R.F., Beiden, S.V., Metz, C.E., “Continuous versus categorical data for ROC analysis: some quantitative considerations,” Acad. Radiol. 8:328–334 (2001).

[27] Beiden, S.V., Wagner, R.F., Doi, K., Nishikawa, R.M., Freedman, M., Lo, S-C B., Xu, X-W., “Independent *versus* sequential reading in ROC studies of computer-assist modalities: analysis of components of variance,” Acad. Radiol. 9:1036–1043 (2002).

[28] Obuchowski, N.A., Beiden, S.V., Berbaum, K.S., Hillis, S.L., Ishwaran, H., Song, H.H., Wagner, R.F., "Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods," Acad. Radiol. 11:980–995 (2004).
[29] Wagner, R.F., Beiden, S.V., Campbell, G., Metz, C.E., Sacks, W.M., "Assessment of medical imaging and computer-assist systems: lessons from recent experience," Acad. Radiol. 9:1264–1277 (2002).
[30] Wagner, R.F., Beiden, S.V., Campbell, G., Metz, C.E., Sacks, W.M., "Contemporary issues for experimental design in assessment of medical imaging and computer-assist systems," Proc. SPIE 5034:213–224 (2003).
[31] Wagner, R.F., Metz, C.E., and Campbell, G., "Assessment of medical imaging systems and computer aids: a tutorial review," Acad. Radiol. 14:723–748 (2007).
[32] Wagner, R.F., "ROC Overview." Tutorial presented at the FDA Radiological Devices Advisory Panel Meeting, March 5, 2001. Available online at: http://www.fda.gov/ohrms/dockets/ac/01/briefing/3734b1.htm Accessed January 30, 2009.
[33] Wagner, R.F., "An overview of contemporary ROC methodology in medical imaging and computer-assist modalities." Tutorial presented at the FDA Radiological Devices Advisory Panel Meeting, February 3, 2004. Available online at: http://www.fda.gov/ohrms/dockets/ac/04/slides/4024s1.htm Accessed January 30, 2009.
[34] Wagner, R.F., Brown, D.G., Guedon, J.-P., Myers, K.J., Wear, K.A., "On combining a few diagnostic tests or features," Proc. SPIE 2167:503-512 (1994).
[35] Wagner, R.F., Chan, H-P, Sahiner, B., Petrick, N., Mossoba, J.T., "Finite-sample effects and resampling plans: Applications to linear classifiers in computer-aided diagnosis," Proc. SPIE 3034:467-477 (1997).
[36] Beiden, S.V., Maloof, M.A., and Wagner, R.F., "A general model for finite-sample effects in training and testing of competing classifiers," IEEE Trans. Pattern Anal. Machine Intell. 25:1561-1569 (2003).
[37] Wagner, R. F., "From medical images to multiple-biomarker microarrays," Invited submission to Vision 2020 series in Med. Phys. 34(12):4944-4951 (2007).
[38] Wagner, R.F., Yousef, W.A., and Chen, W., "Finite training of radiologists and statistical learning machines: parallel lessons," in Advances in Medical Physics, Edited by A.B. Wolbarst, K.L. Mossman, and W.R. Hendee, Medical Physics Publishing, Chapter 9, 129-140 (2008).
[39] Wagner, R.F., "Uncertainties in the multiple-biomarker classifier problem," 7th Meeting of the MicroArray Quality Control (MAQC) Project, titled "Development and Validation of Predictive Models," Cary, NC (May 24-25, 2007).
[40] Chan, H-P, Sahiner, B., Wagner, R.F., Petrick, N., "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Med. Phys. 26:2654-2668 (1999).
[41] Vapnik, V. N., [The Nature of Statistical Learning Theory], Springer-Verlag, New York (1995).
[42] Yousef, W.A., Wagner, R.F., and Loew, M.H., "Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier," Pattern Recog. Letters 26:2600-2610 (2005).
[43] Yousef, W.A., Wagner, R.F., and Loew, M.H., "Assessing classifiers from two independent data sets using ROC analysis: A nonparametric approach," IEEE Trans. Pattern Anal. Machine Intell. 28:1809-1817 (2006).