Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Special Section on Pioneers in Medical Imaging: Honoring the Memory of Robert F. Wagner

Kyle J. Myers
Weijie Chen

SPIE.

Special Section on Pioneers in Medical Imaging: Honoring the Memory of Robert F. Wagner

Kyle J. Myers
Food and Drug Administration
Center for Devices and Radiological Health
Office of Science and Engineering Laboratories
Division of Imaging, Diagnostics, and Software Reliability
E-mail: kyle.myers@fda.hhs.gov

Weijie Chen
Food and Drug Administration
Center for Devices and Radiological Health
Office of Science and Engineering Laboratories
Division of Imaging, Diagnostics, and Software Reliability
E-mail: weijie.chen@fda.hhs.gov

Robert F. Wagner, an SPIE Fellow noted for his achievements in medical imaging, died on June 30, 2008. He was 70.

Bob Wagner, as he was widely known, was a distinguished research physicist and member of the Senior Biomedical Research Service in the Center for Devices and Radiological Health (CDRH), U.S. Food and Drug Administration (FDA). His career was dedicated to the development of consensus measurement methods for the assessment of medical imaging systems, quantitative medical imaging and tissue characterization, and computer-aided diagnosis. He is also remembered for the many invited presentations and tutorials he gave in and outside the FDA, his numerous publications, his professional society activities, his assistance in regulatory decision-making, and his role as a mentor to numerous PhD students, post-docs, and coworkers.

Bob Wagner received his BS in electrical engineering from Villanova University, where he was selected "Outstanding Graduate." He earned an MA in theology from Augustinian College in Washington, D.C., and an MS and PhD in physics from The Catholic University of America. After graduate and post-graduate work on the physics of nuclear interactions with radiation, he was hired by the Bureau of Radiological Health (a precursor to CDRH) to assess the dose reduction potential of radiographic intensifying screens made with phosphors developed in the color TV industry. In 1976 he was named chief of the Diagnostic Imaging Section, and he served in that capacity until 1995, when he assumed the role of FDA Senior Biomedical Research Scientist (SBRS), a position he held until his death.

A Fellow of SPIE since 1988, Bob Wagner was also a fellow of the Institute of Electrical and Electronics Engineers (IEEE), The Optical Society (OSA), American Institute for Medical and Biological Engineering (AIMBE), and Society of Professionals, Scientists and Engineers (SPSE). The FDA honored him with the FDA Commendable Service Award, the Award of Merit, the Commissioner's Special Citation, the Public Health Service Superior Service Award, and the Excellence in Analytical Science Award, presented in 2001 "for the development of multivariate models and software for the assessment of diagnostic tests, imaging, and computer-aided diagnosis in the presence of multiple random effects."

In recognition of his leadership in the field of assessment of diagnostic imaging performance, Bob Wagner was chosen as a principal author of an International Commission for Radiation Units and Measurements (ICRU) report on image quality in medical imaging. The resulting document was published by the ICRU during the centenary year (1995) of the discovery of x rays by Roentgen. This document laid the foundation for a series of ICRU reports with more detailed recipes, one medical imaging modality at a time, that have been developed since.

Bob Wagner served on numerous academic advisory boards, search committees, conference program committees, and editorial boards. He was a prolific reviewer for a broad spectrum of journals, including *Medical Physics, Physics in Medicine & Biology, Optical Engineering*, and *the Journal of the Acoustical Society of America*, and he performed grant review activities for such institutions as the National Science Foundation, the National Cancer Institute, the Atomic Energy Commission, the Canadian Research Councils and those of Great Britain, and the National Institute of Dental Research. In 2001 he co-chaired the annual conference of the Medical Imaging Perception Society.

The papers in this special section are remarkably broad in topic and application area, in line with the many aspects of medical imaging influenced by Bob Wagner's work, including imaging physics, image reconstruction and computer-aided diagnosis, model observers for the prediction of ideal or human performance, and the overarching theme of statistical assessment methodologies for image quality evaluation. Bob Wagner was trained as a physicist and hired by the Bureau of Radiological Health (BRH) to address questions regarding radiation utilization in medical imaging procedures. He quickly formulated a risk-benefit approach to his work, suggesting that the dose associated with the creation of a medical image needed to be considered in light of the usefulness of that image. His earliest SPIE paper was presented in November 1972 in Chicago at SPIE's first dedicated medical imaging meeting, Application of Optical Instrumentation in Medicine. Bob's manuscript, reprinted in this special section, provided an insightful review of the image quantification field, including modulation transfer function (MTF), noise power spectrum (NPS), and receiver operating characteristic (ROC) curves, and a bold statement that laid the foundation for the entire field of medical imaging assessment to follow, that image quality "must be defined in terms of the task that the image is destined to perform."[1]

Several papers address the assessment of image quality for new image acquisition and reconstruction methods, harking back to Bob's earliest work from the 1970s on image assessment in general and the evaluation of new methods for radiography, mammography, and CT in particular. Sidky et al. tackle the problem of the high computational burden for image reconstruction in CT through direct region-of-interest (ROI) image reconstruction. The proposed method is demonstrated for both complete field-of-view and ROI imaging, with applications to actual CT scanner data. Sanchez et al. compare several approaches to estimation of Hotelling observer (HO) performance in x-ray computed tomography (CT). The authors consider the case of signals confined to small regions of interest, enabling direct computation of HO metrics thanks to a reduced dimensionality of the image covariance matrix. Because their method computes HO performance exactly within the ROI, the authors are able to investigate the validity of the assumptions inherent in various common approaches to HO estimation, such as the stationarity assumption often made in Fourier-space analyses. Berglund et al. evaluate energy weighting on a spectral photon-counting mammography system using computer simulations, phantom experiments, and the analysis of screening mammograms. The authors demonstrate the potential for dose reduction for these systems, a modern validation of the theoretical analysis of the advantage of this technology presented by Tapiovaara and Wagner in 1985.[2] Zürch et al. present a new approach to cancer cell classification that makes use of the diffraction pattern of a single cell illuminated with coherent extreme ultraviolet (XUV) laser-generated radiation. These patterns allow distinguishing different breast cancer cell types in a subsequent step. In a proof-of-principle experiment, the authors present data from single breast cancer cells on gold-coated silica slides. Using the resulting diffraction patterns, the authors present evidence for their ability to identify different breast cancer cell expressions.

It is especially fitting that we include a number of papers presenting advances in the area of receiver operating characteristic (ROC) methodology and multireader multicase (MRMC) reader studies for the evaluation of medical imaging systems.[3,4] Bob was an enthusiastic proponent and pioneer in the development of ROC analysis for the assessment of imaging systems, a research effort he led within the imaging group at the FDA for decades, where he mentored and inspired many young scholars during his career who are carrying his torch forward today. In fact, there are four papers in this special section from the imaging group at the FDA where Bob devoted nearly his whole career. These papers include a number of important topics in the areas of ROC modeling, generalization of ROC methodologies to estimation tasks, generalization of MRMC ROC simulation models, and an extension of MRMC methodologies to binary data.

Samuelson and He compare semiparametric ROC models for fitting reader study data and find that the single-parameter power-law model fits many reader study datasets better than two-parameter models (such as the conventional and proper binormal models) in terms of the Akaike and Bayesian information criteria and cross-validation. The findings in this paper suggest that one may give a second thought when fitting the ROC data with a model, i.e., a parsimonious model may be more appropriate when only a dataset of limited size is available.

Wunderlich and Goossens provide practical statistical tools for the evaluation of medical imaging systems in combined detection/estimation tasks, i.e., the task is not only the detection of a signal (e.g., tumor) but rather includes both detection and estimation of a parameter of interest (e.g., tumor size). To evaluate the performance of combined detection/estimation, the notion of an estimation ROC (EROC) curve has been proposed based on the ROC concept for performance evaluation of the signal detection task, but a practical method for estimating the performance figure of merit in EROC has been lacking. Wunderlich and Goossens fill this gap by applying nonparametric statistical techniques to the estimation and statistical inference of the area under the EROC curve.

Gallas and Hillis generalize the Roe and Metz model for simulating decision scores in MRMC ROC studies by explicitly allowing variances of ROC ratings that depend on modality and truth state; such flexibilities may allow more realistic simulations. Furthermore, the analytic link between simulated decision scores and empirical AUC variances and covariances given in this paper may facilitate users in choosing parameters in the simulation model to yield desired ROC parameters. The methodology and the software tools provided by the authors will be useful for investigators for validation of analysis methods in MRMC ROC studies.

Motivated by emerging "whole slide imaging" digital pathology reader studies, Chen et al. extend the MRMC reader study methodology to situations where binary agreement is the study endpoint. These authors developed a statistical model to simulate binary MRMC reader study data in which variability comes from the random reader sample, the random cases, and the interactions thereof. Moreover, they adapted an analysis method that was originally developed for analyzing MRMC ROC data to analyze binary MRMC data. The authors further validated the adapted analysis method using their simulation model and illustrate how to use their simulation model to size a new study.

Bob Wagner also had a strong research interest in the evaluation of computer-aided diagnosis systems and other high-dimensional medical diagnostic classifiers such as DNA microarrays, where the "reader" of the images/microarrays is a computer algorithm.[5] One particular challenge for training and testing computerized classifiers in medicine is the limited patient sample size. Because of this limitation, two important issues arise, namely, (i) the interplay of classifier performance and its uncertainty with sample size and dimensionality (i.e., the number of features),[6] which is vividly depicted by what Bob called the "antler plot" (see Fig. 1), and (ii) the stability of a classifier with respect to varying training datasets, for which Bob Wagner pioneered the notion of "training variability" analogous to the "reader variability" in MRMC reader studies.[7,8]

There are two interesting CAD papers presented in this special section for diagnosis of Alzheimer's disease and breast cancer risk prediction, respectively. Both are exploratory studies with limited datasets using cross-validation for classifier training and validation. Martinez-Torteya et al. used a computerized algorithm to rank the relative contributions of biological, clinical PET, and MRI-related features in a logistic regression model for distinguishing between mild cognitive impairment and Alzheimer's disease progression. This study involved a large number of features and a moderate

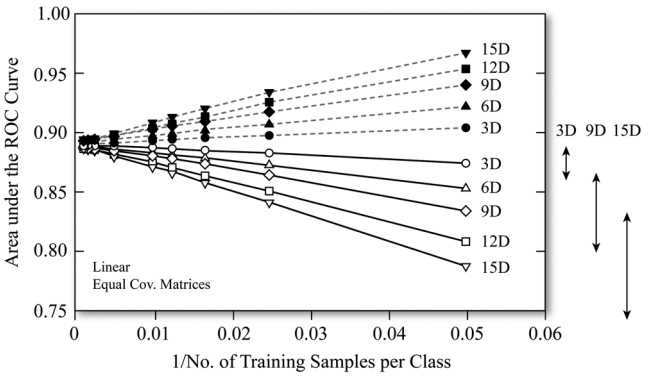

Figure 1 One of Bob Wagner's favorite plots, which he called the "antler plot," depicting the interplay of classifier performance (here AUC on the ordinate axis) with dimensionality (varying from 3D to 15D in this illustrative example) and training sample size (abscissa axis) under two assessment paradigms: resubstitution (training and testing using the same dataset, dashed lines) and independent testing (solid lines). (Adopted from Ref. 7 with permission from Medical Physics Publishing.)

sample size. Using cross-validation, they observed a promising AUC value of 0.79 using the selected features; however, this encouraging point estimate of performance is associated with a large uncertainty (95% CI [0.4, 1] covering the random guess AUC value of 0.5). Note that when not fully cross-validated, an AUC of 0.99 was obtained – an example of the "antler plot!" As the authors suggest, a larger study is needed to further confirm and validate their findings. Li et al. show that parenchymal patterns as characterized by radiographic texture analysis of full-field digital mammograms are promising in distinguishing between high and low risk of breast cancer. The results from their dataset also indicate, surprisingly, that breast density does not appear to be a good risk predictor despite the fact that breast density is a widely recognized risk factor for breast cancer (see references cited by the authors). This is likely due to the limited dataset and the authors are looking forward to expanding their datasets in future studies that would allow for controlling more confounding factors such as menopausal status and hormone replacement therapy status.

A third CAD paper in this special section considers the impact of lesion segmentation metrics on CAD in breast CT. The paper by Kuo et al. compares two segmentation evaluation methods: (i) a Dice similarity coefficient (DSC) evaluation which compares machine segmentations to expert delineations and (ii) a method that takes into account the ultimate performance of the CAD algorithm in the task of classifying malignant from benign lesions in breast CT. The authors conclude that the DSC metric alone is not sufficient for evaluating segmentation lesions in computer-aided diagnosis tasks. This paper, like all the others in this special section, reminds us of Bob Wagner's assertion that the rigorous and objective assessment of imaging systems and algorithms demands the consideration of the ultimate task for which the images will be utilized.

References

1. R. F. Wagner and K. E. Weaver, "An assortment of image quality indexes for radiographic film-screen combinations—can they be resolved?," *Proc. SPIE* **0035**, 83–95 (1972).
2. M. J. Tapiovaar and R. F. Wagner, "SNR and DQE analysis of broad spectrum x-ray imaging," *Phys. Med. Biol.* **30**, 519–529 (1985).
3. R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: a tutorial review," *Acad. Radiol.* **14**(6), 723–748 (2007).
4. S. V. Beiden, R. F. Wagner, and G. Campbell, "Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis," *Acad. Radiol.* **7**, 341–349 (2000).
5. R. F. Wagner, "From medical images to multiple-biomarker microarrays," *Med Phys.* **34**(12), 4944–4951 (2007).
6. H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "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).
7. R. F. Wagner, W. A. Yousef, and W. Chen, "Finite training of radiologists and statistical learning machines: parallel lessons," Book Chapter in *Advances in Medical Physics 2008*, A. B. Wolbarst, K. L. Mossman, and W. R. Hendee, Eds., Medical Physics Publishing, Madison, WI (2008).
8. W. A. Yousef, R. F. Wagner, and M. H. Loew, "Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier," *Pattern Recog. Lett.* **26**, 2600–2610 (2005).