ROC Analysis:
A tribute to Charlie Metz and an assessment of the

State of the Art

Brandon Gallas
FDA/CDRH/OSEL Division of Imaging and Applied Mathematics

SPIE Workshop, 2/10/13
Image Perception, Observer Performance, and Technology

Kyle's comments at memorial

Charles Metz and the FDA had a relationship dating back 40 years. Charlie and Bob Wagner, our dear departed colleague and another giant in the medical imaging field, became acquainted through professional meetings in the 1970s. They found that they shared many things: a passion for science, classical training and far-ranging Renaissance-man intellects, and a love of long, late night phone chats that sent loud laughter through the walls to my office, which was adjacent to Bob's. They tried ideas out on each other, published numerous joint papers, and revolutionized the culture regarding imaging technology assessment in this country.

Charlie, Bob, and Dave Brown, who is here today, first published a paper together in 1981. It was a pioneering paper laying out a task-based approach to comparing gamma-ray imaging systems. I was in grad school at the time, working with Harry Barrett, and this paper changed my life. Just the first of many papers of Charlie's that we all know well, like fine books or poetry, memorizing every detail. That 1981 paper introduced me to concepts I'd never seen before having to do with objective assessments of imaging systems drawn from signal detection theory and applied to medical imaging systems. My work, like so many of you here today, continues to build on the fundamental concepts laid out in that influential early publication.

In the late 1980s, when the ICRU decided to sponsor a monograph on The Assessment of Medical Imaging, they turned to Charlie, Bob, and Dave to form the core of the authors of this report. The timing was masterful – the report was to be released in the year of the 100th anniversary of Roentgen's discovery of x-rays. We at the FDA made great use of that report as unofficial guidance for industry for years.

A landmark moment in regulatory history in this country came on December 11, 1995, when an ROC curve first formed the basis of an FDA advisory panel's deliberations of the benefits and risks of a novel imaging device. The device was an ultrasound imager designed to distinguish benign from malignant lesions that were found indeterminate by mammography and physical examination. At that time, over 700,000 women were undergoing a breast biopsy each year in the US. Up to 80% of those breast lumps would be found to be benign. Charlie worked with the company to design and analyze an international multi-center study involving over 1,000 women, resulting in the finding that the number of breast biopsies could be reduced by approximately 40% using the new device. The FDA Advisory Panel voted unanimously to recommend its approval. It was a great day for women's health, a major milestone at the FDA, and Charlie was at the center of it all.

Then came the approval of digital mammography for breast cancer screening, using Charlie's multi-reader multi-case analyses to demonstrate that digital and film mammography were substantially equivalent. Another major milestone for women's health.

Meanwhile, Charlie's ROC software packages were growing in capability and wide-spread use. For us at the FDA, these publically available tools were extremely important, a gold standard to which we sent countless investigators. They meant that no one had an excuse for not using ROC analysis.

Bob Wagner and I chaired the Medical Image Perception Society in 2001. The conference was scheduled to open on September 20th, and we all know what happened just 9 days earlier. With the airports in Washington closed, and no clear information on when they would reopen, Bob and I contacted key people in the field to discuss our options. Charlie was clear and unequivocal, leading the charge for holding the meeting in spite of the uncertainties. He promised to drive from Chicago. That was that. It was a historic meeting, setting in motion a better paradigm for the evaluation of computer-aided diagnosis systems based on real performance benefits, not potential ones. From that time forward, Charlie was at the forefront of the science around reader studies for CAD.

For many years, a highlight of being in our little FDA group was the annual visits Charlie would pay to us, where he would spend days hearing updates of our research efforts and giving us his piercing feedback. The evenings found us at different restaurants, enjoying his stories and his big laugh. When he cut back on those travels, we missed his personal presence fiercely.

Charlie's influence at the FDA was immense, directly translatable to millions of patients getting earlier access to effective medical imaging devices and CAD systems. For those of us at the FDA who had the privilege to work with him, he was a champion of solid scientific methods and data, a source of inspiration personally and professionally, and a dear friend and colleague.

Bob Wagner

- Charlie and Bob
  - "fast friends" … "bromance"
  - Late night marathon phone calls
  - Lots of laughing
- 1970's and first SPIE
- 1981 first paper together (incl. Dave Brown)
  - Task-based Assessment takes root in medical imaging
  - Credit to Lusted, Green, Swets …
  - Likelihood ratio

the FDA

- Charlie had a major impact on FDA business
  - like-minded FDA scientists: The Three Bobs (Wagner, Jennings, Gagne), Dave Brawn, Mary Pastel, Kyle Myers
- ICRU: The Assessment of Medical Imaging
  - work started in 1980's, published 1995
- ROC: first used to support an imaging device approval (1995)
  - Charlie did the study design and analysis
- DBM: MRMC analysis first used to demonstrate digital mammography substantially equivalent to film

  DBM = Dorfman, Berbaum, Metz method

  MRMC = Multi-reader multi-case analyses software

  (LABMRMC software hosted at U. Chicago)

the FDA/MIPS

- MIPS 2001: set in motion paradigm to evaluate CAD with human in the loop
  - days after 9/11, he promised to drive
  - MIPS = Medical Image Perception Society
- FDA/MIPS 2010: Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA
  - Workshop Summary, Acad Radiol, 2012
- MIPS 2013: August, George Washington, DC
  - FDA centric Friday
  - reader studies to support imaging device approvals
  - study designs and analysis methods

Heang-Ping's Grant
Research updates: U Michigan, Charlie, Bob

Software

- http://metz-roc.uchicago.edu/
- 

- http://js.cx/~xin/index.html
- iMRMC Analyzing and Sizing Multi-reader M
case ROC Trials

Xin He, Brandon Gallas

Incl. Database of Components of Variance

I want your reader data!

1-degree of Separation!

- Gallas, B. D.; Chan, H.; D'Orsi, C. J.; Dodd, L. E.; Giger, M. L.; Gur, D.; Krupinski, E. A.; Metz, C. E.; Myers, K. J.; Obuchowski, N. A.; Sahiner, B.; Toledano, A. Y. & Zuley, M. L. (2012),
- 'Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA.'
- *Acad Radiol* ***19(4), 463-477.***

- "This is an *excellent* questionaire overall, in my opinion."
- "This document is very, VERY good, in my opinion."
- "This sentence strikes me as not only awkward, but impenetrably opaque."

AAPM 45th Annual Meeting, 2003

- **Practical Aspects of CAD Research**
  - Assessment Methodologies for CAD
  - http://www.aapm.org/meetings/03AM/pdf/9851-59652.pdf
- **Models in Medicine IV**
  - ROC Analysis: Methods and Practical Applications
  - http://www.aapm.org/meetings/03AM/pdf/9850-28518.pdf
- **"Incomplete" list of recommended literature on ROC methodology**
  - 86 references: Background, General, Bias, Curve Fitting, Statistics, Relationships with Cost/Benefit Analysis, Generalizations

AUC vs. …

- Interpretations of ROC area …
- However, …
  - this global index can be misleading when curves cross and/or there is only one region of interest
- Other ROC-based indices of performance
  - Partial area below, or to the right of, a segment of the ROC curve (regional)
  - TPF at fixed FPF or vice-versa (local)
  - Expected utility at optimal operating point (local) — most meaningful but least practical

- WS Questionnaire: "In principle, utilities most certainly should be defined and incorporated in a review. In practice now and for the foreseeable future, this is impractical and should be avoided."

Expected Utility

- "Statistical Power Considerations for a Utility Endpoint in Observer Performance Studies."
  - Abbey, Samuelson, and Gallas
  - Academic Radiology Special Issue for Metz (accepted)
  - "expected utility (EU) endpoint that is based on the observed relative utility of screening mammography"
  - Given DMIST and BCSC
  - EU endpoint generally has good statistical power relative to AUC in our simulations

Needs for the future

- Charlie's slide
- Possible discussion points

Needs for the future

- Develop stratified-sampling methodology
- Establish validity/robustness of data-analysis techniques for free-response paradigms
  - curve fitting
  - statistical testing of differences
- Develop "MRMC" methods for statistical analysis of data from incompletely-balanced experimental designs, particularly ...
  - when observers don't read the same cases
  - when data are correlated within cases
- Develop highly efficient approaches well-suited to exploratory analyses
  - Key need is to control for decision-threshold effects
  - Other biases may be acceptable if sufficiently small
- Generalize ROC analysis to handle >2 decision alternatives
  - Must provide an appropriate compromise between complexity and practicality
  - Approaches proposed to date are not adequate

Stratified-Sampling
Example: FFDM in screening

- Prospective Clinical Study
  - Prevalence < 0.5%
  - All screening patients
  - DMIST: 42,760 cases
  - 2 readers per case
  - FFDM vs. SFM
  - FFDM AUC = 0.78 ± 0.02
- Controlled Retrospective Reader Study
  - Enrich: Overweight patients with cancer
  - Stress: Overweight recalled patients
  - Hologic: 312 cases
  - 12 readers (fully crossed)
  - FFDM vs. FFDM+Tomo
  - FFDM AUC: 0.82 ± 0.04

Comparing AUCs

Stratified-Sampling
Example: FFDM in screening

- Prospective Clinical Study
  - Screening recall rate 8.4%
  - Pooling SFM "or" FFDM, recall rate is 14%
  - FFDM AUC = 0.78 ± 0.02
- Controlled Retrospective Reader Study
  - Pooling FFDM "or" FFDM+DBT intake recall rate 26%
  - Study population of normals included 53% recalled
  - FFDM AUC = 0.82 ± 0.04

AUCs are not comparable

Different populations
→ Different performance

Stratified-Sampling
Pinsky and Gallas, Stat. Med. (2012)

- Stratification variables
  - Truth status
  - Recall status
- Overweight by truth status
  - AUC is not biased
- Overweight by recall status
  - AUC is Biased
- Undo bias by inverse probability weighting
  - Pay for it with increased variance
- Bias is a function of
  - Correlation between test and stratification variable
- Direction of bias
  - can be up or down ☹
  - can design the sampling to be against new modality (under reasonable assumptions or, better, knowing correlations)
- Math not human behavior
- Still more work to be done

incompletely-balanced experimental designs

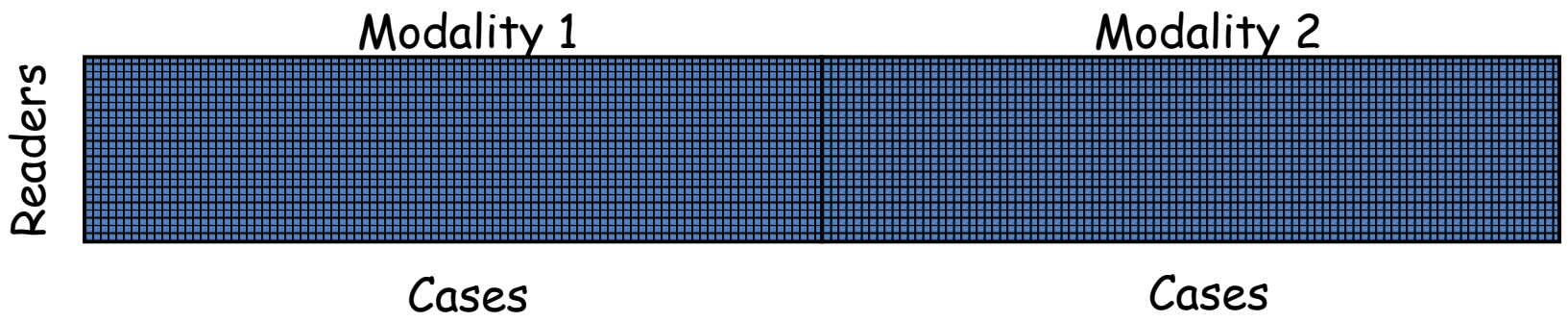
- MRMC variance estimation for arbitrary study designs
  - Binary Performance: Gallas et al. (2007) JOSA
  - AUC: Gallas & Brown (2008), Neural Networks
  - U-statistics
  - A design matrix controls/shows data collected vs. fully-crossed

- 'Multi-Reader ROC Studies with Split-Plot Designs: A Comparison of Statistical Methods'
  - Obuchowski, Gallas, Hillis (2012)
  - Acad Radiol Special Issue for Metz

- Revolutionize Study Designs
  "in my opinion"

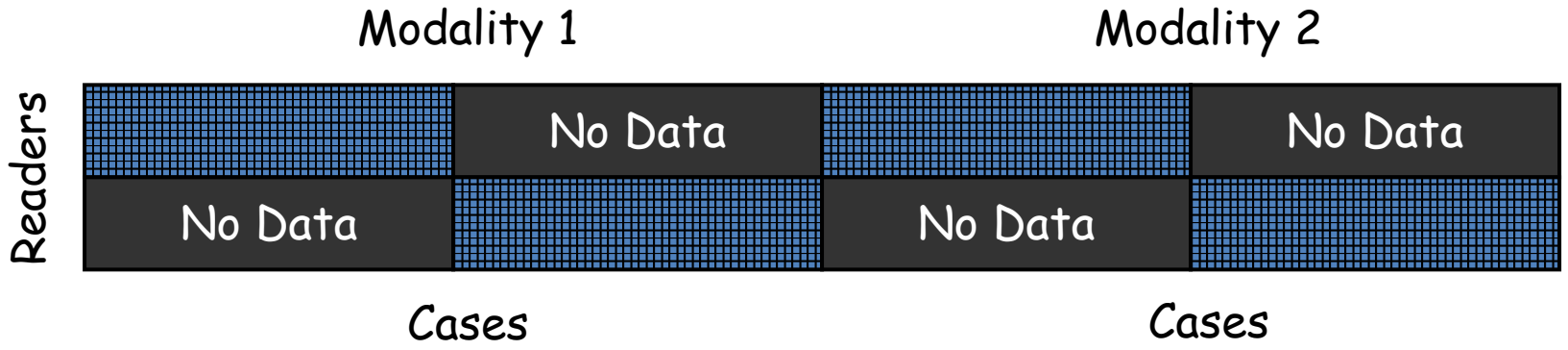
Study Designs

- Fully-crossed study
  - All readers read all cases
  - Readers and cases are paired across modalities



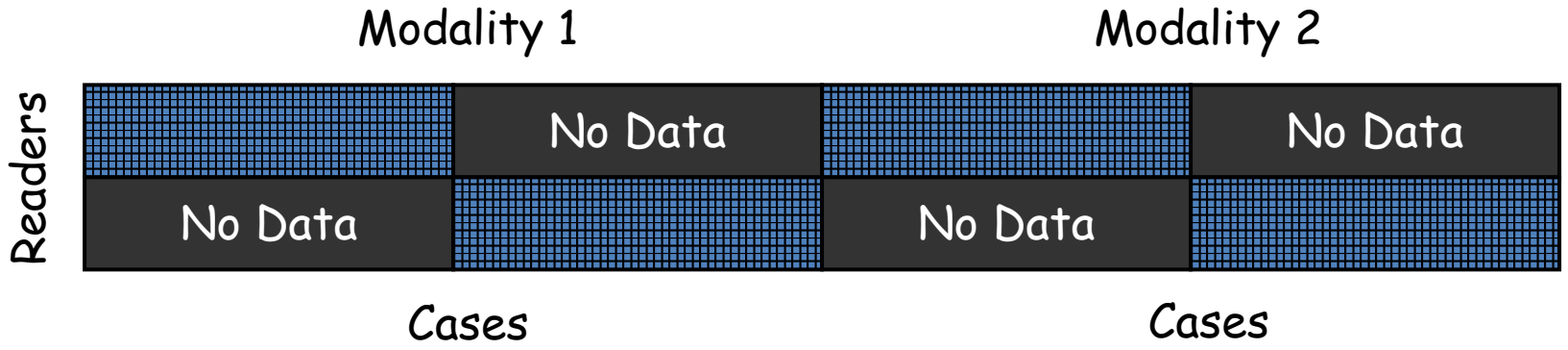
Study Designs

- Fully-crossed study is burdensome
  - All readers read all cases
  - Readers and cases are paired across modalities
- Split-plot study
  - Readers and cases split into 2 groups
  - Data is fully-crossed within a group



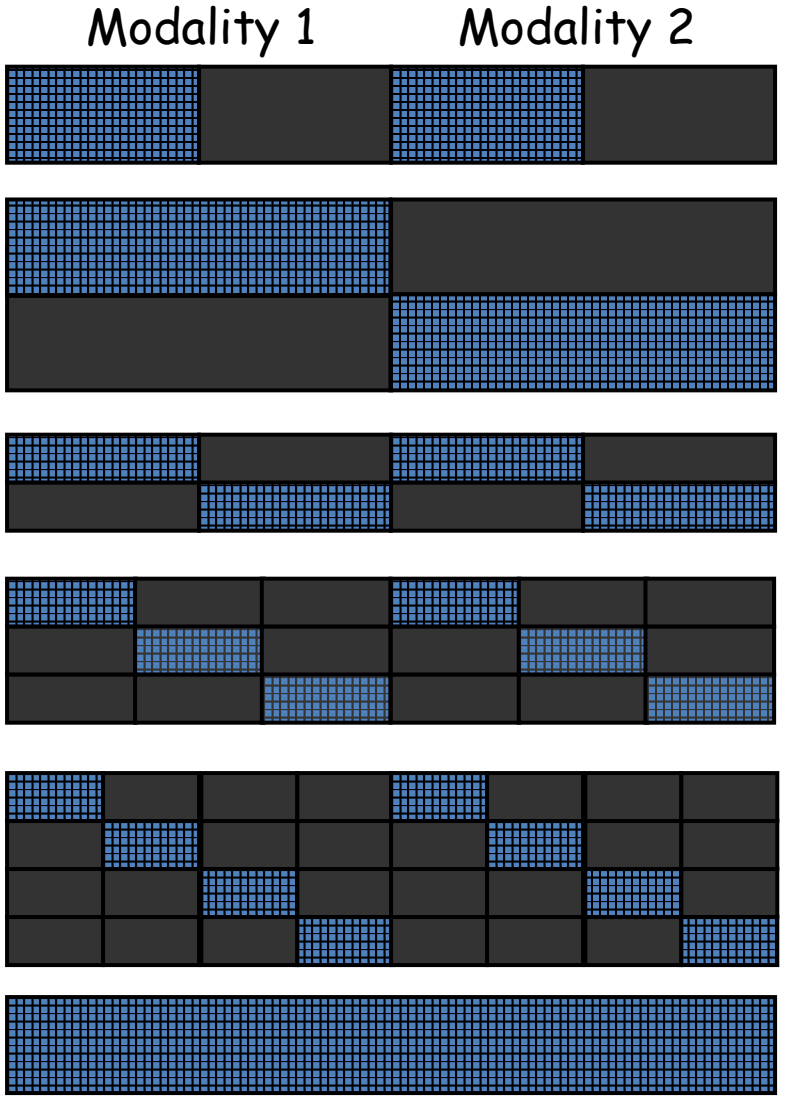
Study Designs

- Fully-crossed is burdensome
  - A lot of reads per reader
  - A lot of reads total
- Split-plot study may save time (and money)
  - Half the reads per reader
  - Half the reads total



Study Designs

- Fully-Crossed A
- Readers Unpaired Across Modalities
- 2-Groups
- 3-Groups
- 4-Groups
- Fully-Crossed B



Study Designs
Efficiency

- Roe and Metz simulation
  - given description of scores, know the components of variance
  - numerical integration

Study Designs: Efficiency

Study Design	Groups	Readers		Cases		Reads		Statistical Efficiency
	G	$\frac{J}{G}$	J	$\frac{N_1}{G}+\frac{N_0}{G}$	N_{Total} N_0+N_1	per reader	total	$\frac{\text{var(2-groups)}}{\text{var(alt. design)}}$
Full-A	1	6	6	30+30	60	120	720	0.83
Unpaired Readers	1	6	12*	60+60	120	120	1440	0.90
2-groups	2	3	6	30+30	120	120	720	1.00
3-groups	3	3	9	20+20	120	80	720	1.20
4-groups	4	3	12	15+15	120	60	720	1.33
Full-B	1	6	6	60+60	120	240	1440	1.16

Resources: Tried to control
- total # reads
- total # cases
- # reads per reader

Study Designs: Efficiency

Study Design	Groups	Readers		Cases		Scores		Statistical Efficiency
	G	$\frac{J}{G}$	J	$\frac{N_1}{G}+\frac{N_0}{G}$	N_{Total} N_0+N_1	per reader	total	$\frac{\text{var(2-groups)}}{\text{var(alt. design)}}$
Full-A	1	6	6	30+30	60	120	720	0.83
Unpaired Readers	1	6	12*	60+60	120	120	1440	0.90
2-groups	2	3	6	30+30	120	120	720	1.00
3-groups	3	3	9	20+20	120	80	720	1.20
4-groups	4	3	12	15+15	120	60	720	1.33
Full-B	1	6	6	60+60	120	240	1440	1.16

Take-away 1. It is possible (and fairly easy) to compare study designs.

Study Designs: Efficiency

Study Design	Groups	Readers		Cases		Scores		Statistical Efficiency
	G	$\frac{J}{G}$	J	$\frac{N_1}{G}+\frac{N_0}{G}$	$\frac{N_{Total}}{N_0+N_1}$	per reader	total	$\frac{\text{var(2-groups)}}{\text{var(alt. design)}}$
Full-A	1	6	6	30+30	60	120	720	0.83
Unpaired Readers	1	6	12*	60+60	120	120	1440	0.90
2-groups	2	3	6	30+30	120	120	720	1.00
3-groups	3	3	9	20+20	120	80	720	1.20
4-groups	4	3	12	15+15	120	60	720	1.33
Full-B	1	6	6	60+60	120	240	1440	1.16

Take-away 2. Pay a price when you don't pair readers across modalities.

Study Designs: Efficiency

Study Design	Groups	Readers		Cases		Scores		Statistical Efficiency
	G	$\frac{J}{G}$	J	$\frac{N_1}{G}+\frac{N_0}{G}$	$\frac{N_{Total}}{N_0+N_1}$	per reader	total	$\frac{\text{var(2-groups)}}{\text{var(alt. design)}}$
Full-A	1	6	6	30+30	60	120	720	0.83
Unpaired Readers	1	6	12*	60+60	120	120	1440	0.90
2-groups	2	3	6	30+30	120	120	720	1.00
3-groups	3	3	9	20+20	120	80	720	1.20
4-groups	4	3	12	15+15	120	60	720	1.33
Full-B	1	6	6	60+60	120	240	1440	1.16

Take-away 3. There is a moderate hit to efficiency when you split the experiment into two groups.

Study Designs: Efficiency

Study Design	Groups	Readers		Cases		Scores		Statistical Efficiency
	G	$\frac{J}{G}$	J	$\frac{N_1}{G}+\frac{N_0}{G}$	N_{Total} N_0+N_1	per reader	total	$\frac{\text{var(2-groups)}}{\text{var(alt. design)}}$
Full-A	1	6	6	30+30	60	120	720	0.83
Unpaired Readers	1	6	12*	60+60	120	120	1440	0.90
2-groups	2	3	6	30+30	120	120	720	1.00
3-groups	3	3	9	20+20	120	80	720	1.20
4-groups	4	3	12	15+15	120	60	720	1.33
Full-B	1	6	6	60+60	120	240	1440	1.16

Take-away 4. You can be more efficient by splitting more. (need more readers, should avoid splitting below 25 cases per truth per reader)

iMRMC 2.0

in development

iMRMC Analyzing and Sizing Multi-reader Multi-case ROC Trials

Xin He, Brandon Gallas

- Allow for arbitrary study design

- Roe and Metz App in development
  - Simulate MRMC experiments
  - Allow variance to differ across truth and modality
  - Numerically calculate components of variance

My Times UP!