Gap Analysis: What Should Be Done to Accelerate QI in Tool Development and Validation?

Anthony P. Reeves
Vision and Image Analysis Group
School of Electrical and Computer Engineering
Cornell University

Lung Cancer Workshop IX, May 3, 2012

Disclosure

Commercial relationships

1. **VisionGate, Inc**.: Dr. Reeves is a paid consultant and holds stock in the company. VisionGate is developing optical imaging technology for the analysis of individual cells.
2. **General Electric**: Dr. Reeves is a co-inventor on a patent and other pending patents owned by Cornell Research Foundation (CRF) which are non-exclusively licensed and related to technology involving computer-aided diagnostic methods, including measurement of nodules

Research Support:

NCI, NSF, American Legacy Foundation, Flight Attendants' Medical Research Institute, AstraZeneca, Inc., and GlaxoSmithKline.

Theme

- 2010 Percolating
- 2011 Brewing
- 2012 ***Boiling***

- Volumetric Analysis does not mean measuring volumes
- Human (expert) intervention is considered harmful

Image Biomarkers

- Classical CAD
  - Disease detection: Evaluate on clinical data FROC analysis
  - Disease diagnosis: Evaluate on clinical data ROC analysis
  - Validation: Do a comparative study with users and obtain a significant p-value

- Quantitative Image Biomarkers
  - Performance depends upon:

    (a) the technical precision of the measurement and

    (b) the clinical efficacy of that measurement
  - **Challenge for lung cancer**: precise ground truth is not known

Pulmonary nodules: ground truth is not known

Phantoms (synthetic nodules with known ground truth)

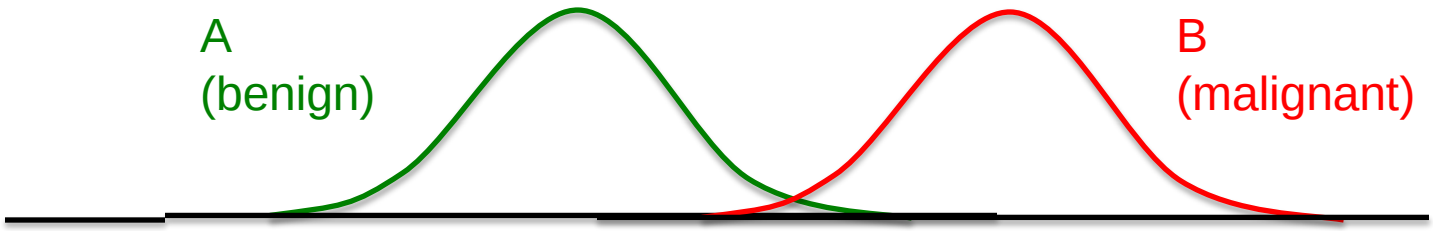
- **Calibrated phantoms are very useful** for calibrating a system or **for testing** that a system is calibrated and operating correctly.
- **Phantoms are not like real data** and only have limited utility in training and validating computer algorithms

Measurements by experts

From the LIDC study for pulmonary nodules:

- There is a very large variation between the “experts”
- We should expect that a good algorithm will be more consistent than the “experts”
- **Issue:** how can we use poor quality ground truth if at all?

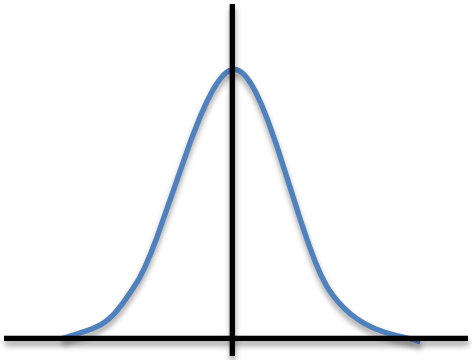
Quantitative Image Biomarkers



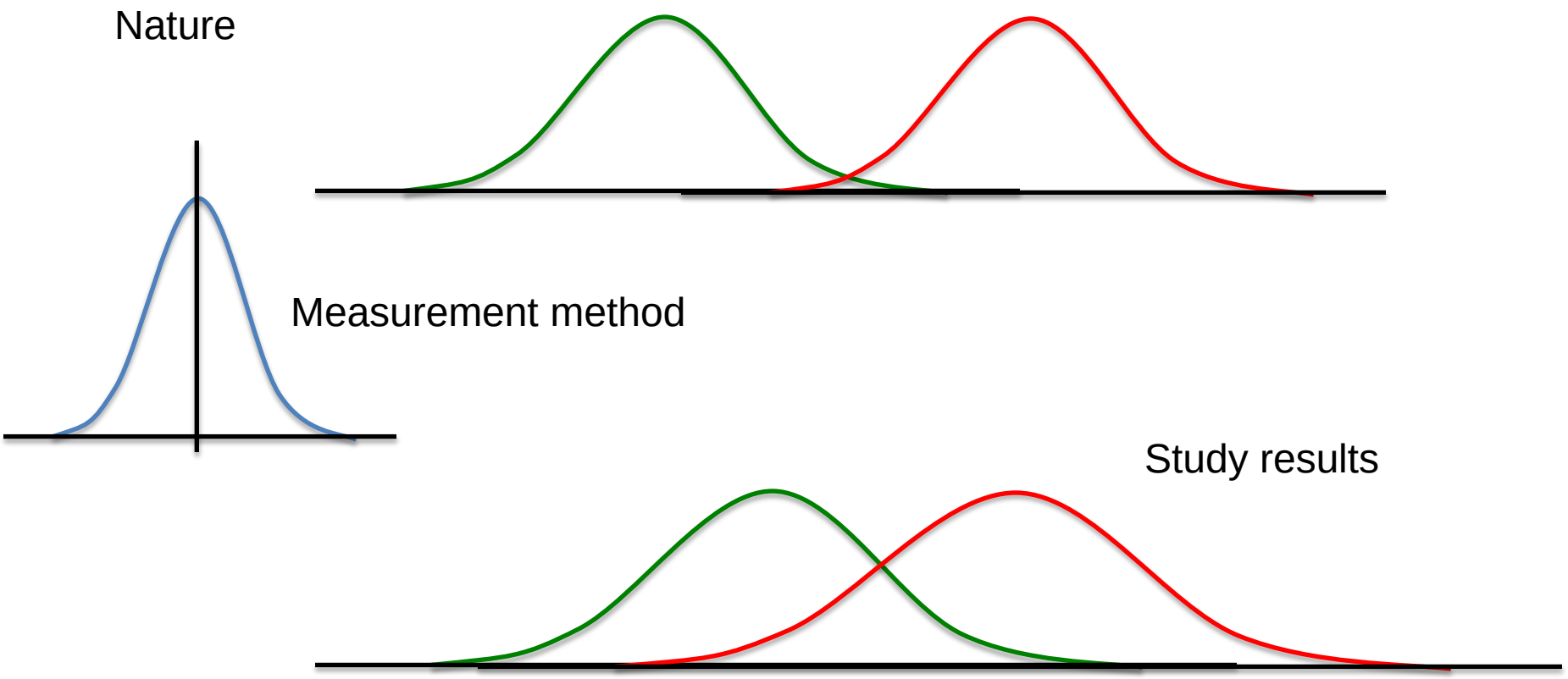
Quantitative Measurement ➔

- **Concept:** a quantitative measurement will distinguish between two medical conditions

- Quantitative measuring devices have uncertainty (variation) associated with their measurements



Quantitative Image Biomarkers



- Observed study outcome is diminished by the variation in the measuring method
- **Issue:** we cannot determine nature (clinical efficacy)

Quantitative Biomarkers

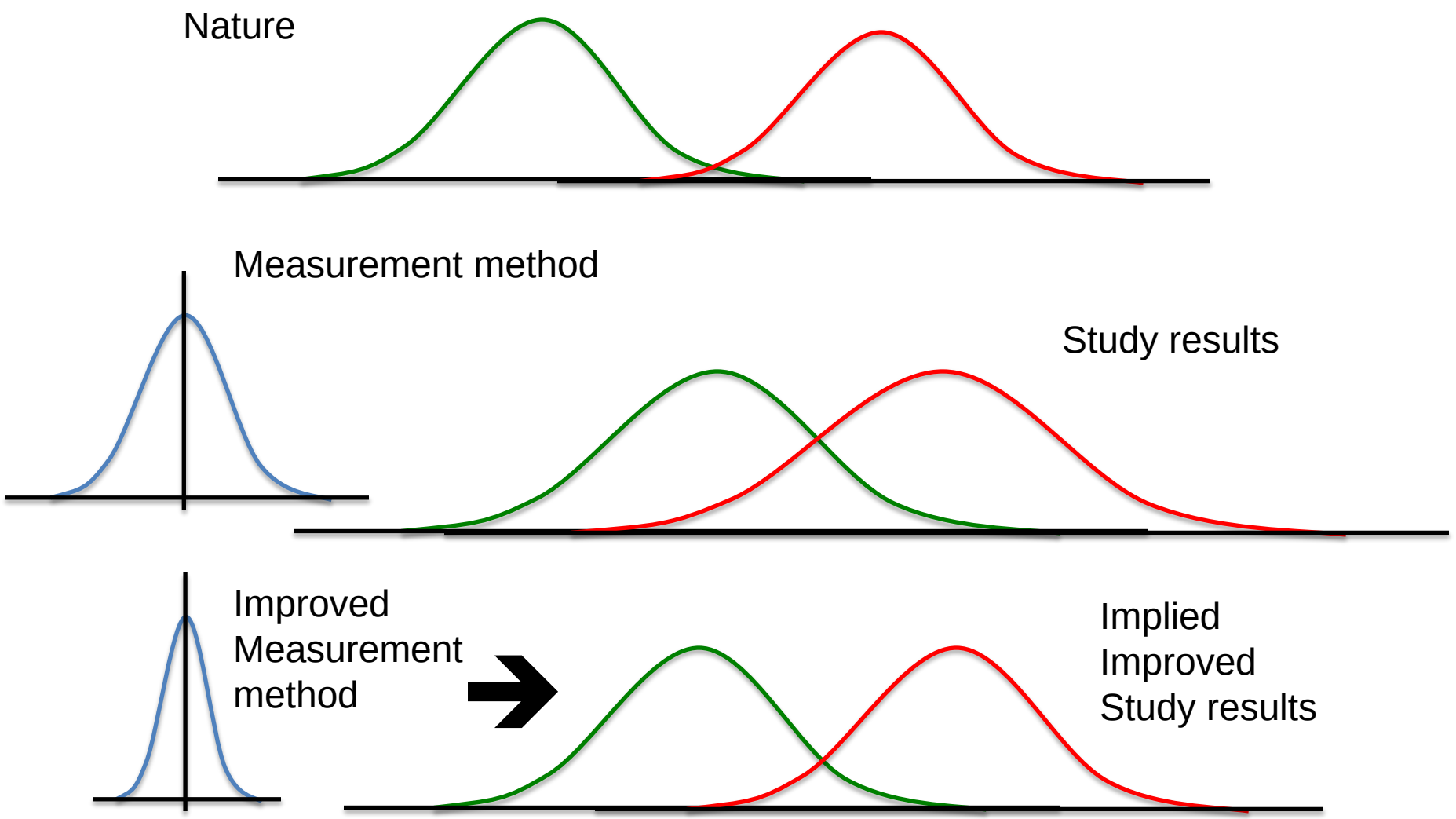
Basic design concept

If the technical performance of the biomarker measurement is improved then the clinical performance of the device is also improved.

---- until the clinical efficacy effect dominates

A new clinical study is not required

Quantitative Objective measurand



Pulmonary nodule measurement in CT images

1. **Response to therapy**

   How much effect did the therapy have on changing the size of the nodule?

   Traditionally, *because of measurement err*or, categorized to: increase, decrease, don't know

   Size range 10-20 mm to > 100 mm

2. **Diagnosis of early stage cancer (small size nodule)**

   What is the growth rate of the nodule?

   High growth rate: cancer, low growth rate: benign.

   Size range < 4 mm to 20 mm

Quantitative Biomarkers (digital)

At my local drug store

- Weight
- Blood Pressure
- Heart Rate
- Blood Glucose
- Temperature
  - Fever
  - Basal
  - Room
  - Outside

Thermometers

	Type	Function	Range °F	Resolution	Accuracy
1	Fever	Fever	90 - 119	0.1	±0.2
2	Basal	Ovulation	89.6 - 109.4	0.01	±0.1 (95-100)
3	Laboratory	Hypothermia	-58 - 308	0.1	±2
4	Room	Environment	23 - 122	0.2	?
5	Infrared	Prevention	-22 - 518	0.1	±5

Thermometers

	Type	Function	Range °F	Resolution	Accuracy
1	Fever	Fever	90 - 119	0.1	±0.2
2	Basal	Ovulation	89.6 - 109.4	0.01	±0.1 (95-100)

Cost $6 to $12

NIST Traceable® Digital Thermometers

–58.000 to 302.000°F 0.001° resolution

Accurate to ±0.02°F

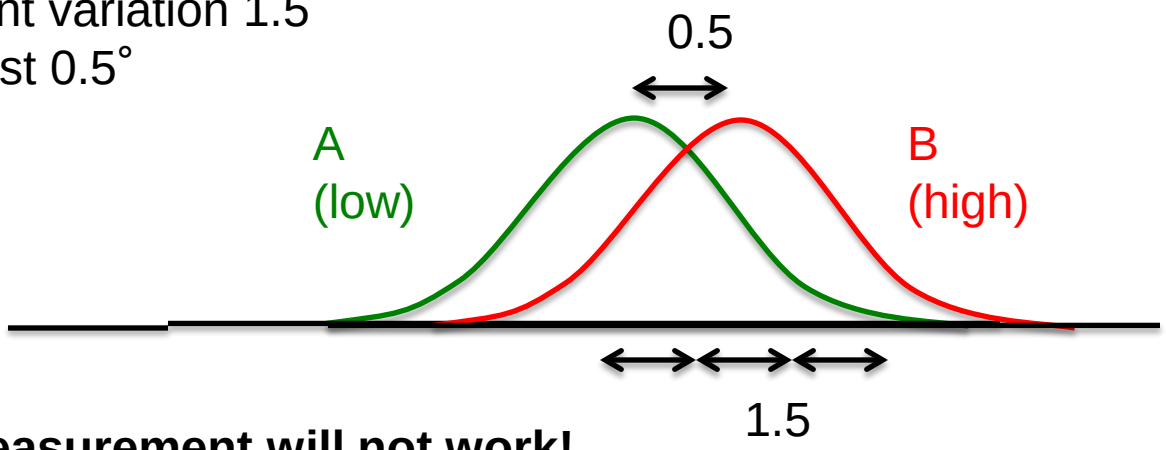
Cost $473

Using a basal thermometer

- Charts to plot the results have a scale of 0.1 or 0.2F
- Multiple readings are made to obtain a reliable outcome
- Although the generally accepted “normal” temperature of a healthy person is 98.6˚ F, the basal oral temperature in the first part of a cycle is usually in a range between 96.50˚ F and 98.00˚ F. In approximately the last two weeks of the cycle, the temperature is typically 0.05˚ F higher.
- Population variation 1.5˚ measurement of interest a change of 0.5˚

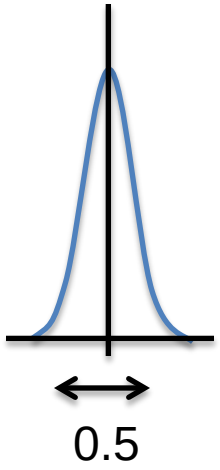
Basal thermometer analysis

Nature
Between patient variation 1.5˚
Effect of interest 0.5˚



Absolute measurement will not work!

- **Design criteria: The repeatability of the measurement should be smaller than 0.5˚**
- The smaller the better until the intra-patient variability dominates
- For intra-patient variability use repeated measures



Quantitative Image Measurement types

V volume, L length, A area, D diameter, I intensity

{Vx} volume occupancy, {Ax} area occupancy, Δt time interval

f(t) function with respect to time

Pulmonary nodule measurement in CT images

1. Response to therapy

   Measurement type: 3: Proportional change (volume)

   Measurement is made on **two** images

   Size range: 4000 mm^3 to > 1000000 mm^3

   Clinically relevant issue: how small can we make the "don't know" category

2. Diagnosis of early stage cancer (small size nodule)

   Measurement type: 4: Growth rate (volume)

   Measurement is made on **two images and Δt**

   Measurement error is inversely related to Δt

   Size range: 300 mm^3 to 4000 mm^3

   Clinically relevant issue: how small can we make Δt

Issue: Measurement Types

For algorithm validation it is important to study the correct measurement type.

[objectively measured characteristic]

Volumetric Analysis does not imply that we measure volumes

Human Interaction Harmful

Human Intervention

Concept: combine the advantages of computer and expert

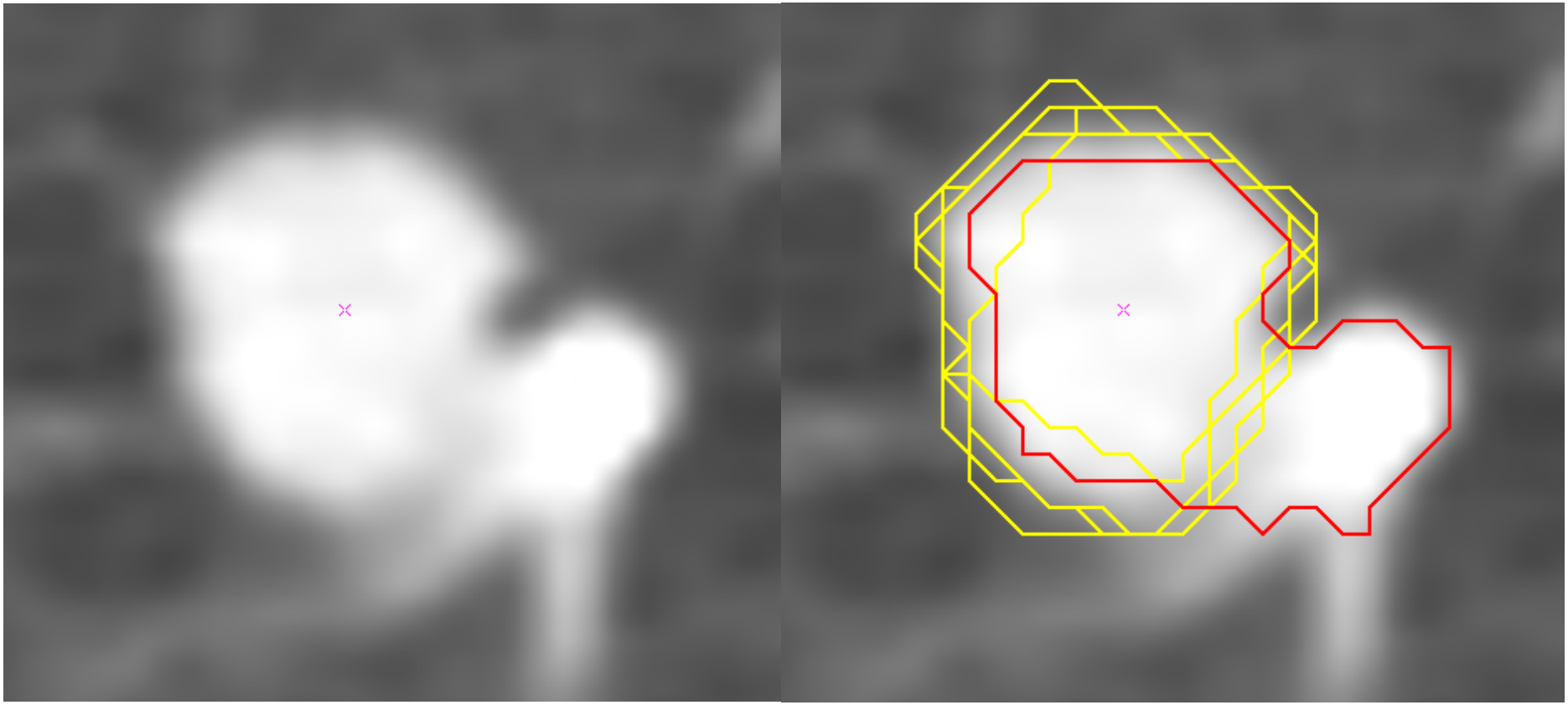
Modes of computer algorithm operation

1. **Active human intervention** (human modifies algorithm response)
2. Automated measurement with **human review** (and outcome rejection)

- Most image measuring computer algorithms permit/require human intervention (especially if they are seeking FDA approval).
- Issue 1: result may combine the **disadvantages** of computer and expert.
- Issue 2: the validation of a human assisted computer algorithm is vastly more complex and more expensive than for a fully automated algorithm.

Example marked nodule (LIDC)

Issue 1: disadvantages of both



Issue: Validation with Human Intervention

Issue 2: cost

With Human Intervention

1. Inter-reader variability
2. Intra-reader variability
3. Study must deal with reader fatigue, memory effects, etc.
   (readers know that patient care does not depend upon these reads)
4. Controlled workstation environment for quality reads
5. ***Cost limits study size to a few hundred cases***

Fully Automated System

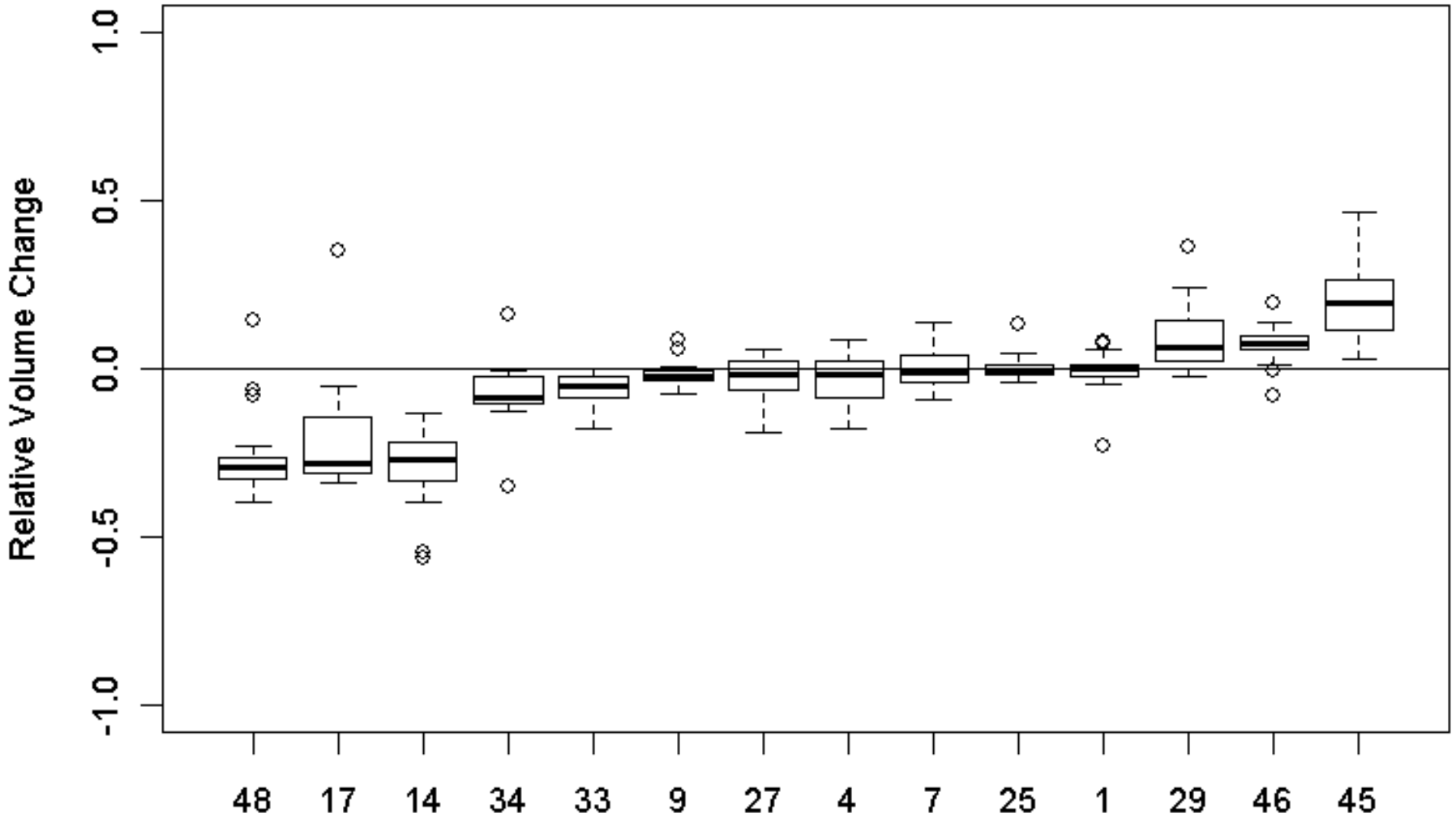
1. Rent cloud computing from Amazon.com
2. Study size not strictly limited (100,000 cases or many more)

Goal: Provide standardized benchmarks results for pulmonary nodule **change** analysis

First benchmark data set: 50 cases of image pairs real change(22), zero-change (27), phantom (1)

1. Study 1: 17 computer algorithms, non-parametric stats.
2. Study 2: Radiologists performance on same data set
3. Provide a public resource of **relevant** cases with **extensive** documentation for algorithm benchmarking

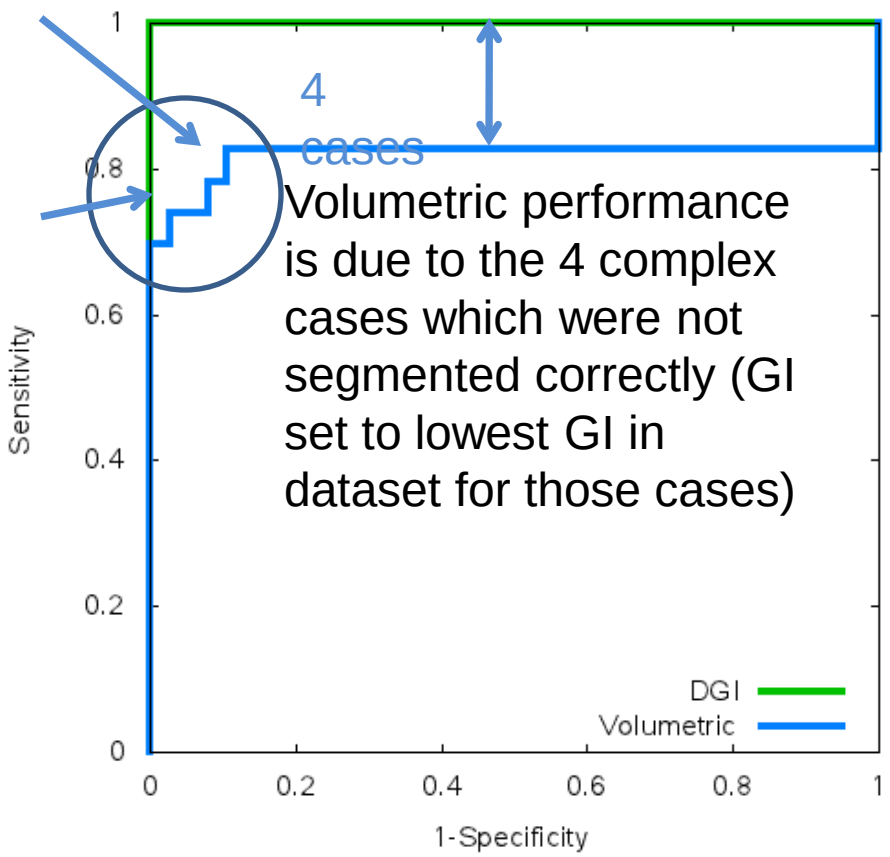
Results: Group A
zero-change, same slice thickness



Diagnostic Performance

Dataset:
38 benign
23 malignant

AUC:
DGI: 1.00
Volumetric: 0.817



Operating point at best sensitivity

Method	Benign	Malignant
Volumetric	90% (34/38)	83% (19/23)
DGI	100% (38/38)	100% (23/23)

ISBI 2009

Validation of lung biomarkers

1. Response to therapy
   - Tools should be validated to a given precision
   - For a sufficiently large documented data set what is the conditional measurement uncertainty (size change?) for no errors (x% errors) with respect to disease progression and response to therapy.
   - Condition: nodule size range (4000 mm^3 to > 1000000 mm^3)

2. Diagnosis from Growth rate
   - For a given device what is the conditional measurement uncertainty for a correct diagnosis to be made (or for x% errors).
   - Condition 1: nodule size range (300 mm^3 to 4000 mm^3)
   - Condition 2: Interval between scans of XX days

Quantitative Biomarker Comparison

	Ovulation	Lung cancer
Device	Basal Thermometer	CT scanner
Design	Specific	General
Protocol	Fixed by manufacturer	? User specification
Quality Assurance	Multiple readings Change device	? User approval
Human interaction	No	Yes
Validated performance	Yes	No

- Publications that do descriptive statistics on doubling times and do not account for measurement error
- The devil is in the details

Summary

1. Quantitative Image Biomarkers should not be validated using methods designed for qualitative biomarkers
   - Less emphasis on p-values more emphasis on confidence intervals of limits of agreement
   - Quantitative Image biomarkers should be developed and evaluated based on objective technical requirements

1. Volumetric analysis does not mean measuring volumes
   - Validation should be made on the ***appropriate*** measurand using ***real data*** not phantom data

2. Human intervention is a major problem with quantitative biomarker validation and should be eliminated.