A framework for utilizing ROC methodology in radiotherapy QA

L Archambault

Assistant professor, Department of Physics, Laval University, Quebec city, QC, Canada
Medical Physicist, CHU de Québec
louis.archambault@phy.ulaval.ca

Plan

- Introduction
- Uses of ROC analysis for QA
  - Compare plan quality metrics
  - Quantify detector performance
  - Improve IMRT/VMAT pre-treatment QA
- Bringing ROC analysis in the clinical routine?

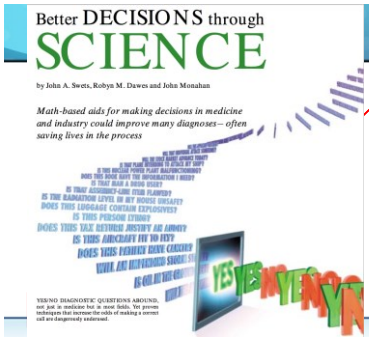
QA: pre-treatment / end-to-end / …

- When performing QA:
  - Multiple detectors can be chosen
    - Diode, ion chambers, films, EPID, gels
    - 1D, 2D, 3D
  - Multiple tests
    - Gamma implementation
    - 2D, 3D
  - Multiple definitions of pass/fail

A large amount of possible combinations, each can yield quite different results

Context

- Is an IBA MatriXX better at catching a single bad MLC leaf using my homemade gamma software with 3 %/ 3 mm than an EPID with a commercial gamma calculation using 2 %/ 2 mm?
- What detector should I use to catch a problem with the penumbra beam model in Eclipse?


Better DECISIONS through
SCIENCE
by John A. Swets, Robyn M. Dawes and John Monahan
Math-based aids for making decisions in medicine and industry could improve many diagnoses—often saving lives in the process
YES/NO DIAGNOSTIC QUESTIONS ABOUND, not just in medicine but in most fields. Yet proven techniques that increase the odds of making a correct call are dangerously underused.

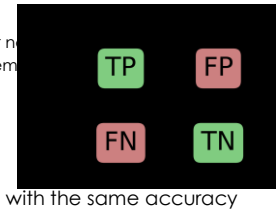
Yes/No diagnostic questions abound, not just in medicine but in most fields. ***Yet proven techniques that increase the odds of making a correct call are dangerously underused.***

We could argue that this is still the case for RT QA

Scientific American, October 2000

QA: pre-treatment / end-to-end / …

- Task of QA: find an error
  - Binary result: is there an error or n
  - This is a signal detection problem
    - The error is the signal



- Systems do not detect errors with the same accuracy

Error detection

- A given detection system:
  - Detector, test, pass/fail threshold
- Capacity to detect error can be characterized
  - **Sensitivity**: fraction of time a positive result is 'real'

$$\boxed{\frac{TP}{TP+FN}}$$

  - **Specificity**: fraction of time a negative result is 'real'

$$\boxed{\frac{TN}{TN+FP}}$$

Why bring ROC curves to RT QA?

- QA methods often have 'knobs' to adjust
  - For example, when using a simple 2D gamma test:
    - Which pixel to consider (e.g. % of prescribed dose)
    - % dose difference (%DD), distance to agreement (DTA)
    - % of pixels that must fail to consider an error
- Different systems have different optimum choices
  - It is unfair to compare different systems using the same %DD, DTA and pass rate
    - Each should use its optimum parameters

Why bring ROC curves to RT QA?

- We believe ROC curves are the answer:
  - Offers an objective framework to compare QA systems
    - Account for detector, test, threshold
  - Easily and visually compare systems independently of the 'knobs' settings
  - Assess how a QA system perform for specific type of error

Uses of ROC analysis in the literature

ROC for QA in the literature

- ROC formalism to improve RT QA is still in its early stage
  - First use (to my knowledge) in 2005

    Childress et al. *Detection of IMRT delivery errors using a quantitative 2D dosimetric verification system*, Med. Phys. 2005
  - Since then: about a dozen papers on the topic

How has ROC been used in RT

- Main applications:
  - Compare plan quality metrics
    - Plan quality, robustness, complexity
  - Quantify detector performance
  - Improve IMRT/VMAT pre-treatment QA
    - Assess the capacity of QA to detect specific type of errors
    - Find the optimal parameters of a test
    - Compare tests

Uses of ROC analysis in the literature

Comparing plan quality metric

Plan quality metrics

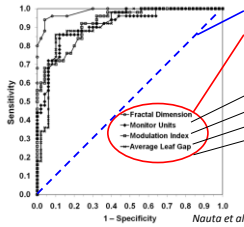
- A number (*metric*) extracted from the plan that is an indicator of the plan complexity and/or quality
  - From TG-119: "*the level of complexity of individual plans is related to the delivery accuracy*"
- Questions
  - How does different metrics compare to one another?
  - Do metrics predict plan quality?

ROC analysis

Some examples

- Nauta et al. Med. Phys. 2011
  - **Mertric**: fractal dimension
  - **Test**: Identify plans with high/low fluence smoothing
- McNiven et al. Med. Phys. 2010:
  - **Metric**: Modulation complexity score (MCS)
  - **Test**: pass/fail of pre-tx QA
- Garcia-Romero et al. Med. Phys. 2016
  - **Metric**: DVH based, robustness, changes in TCP/NTCP
  - **Test**: Dose difference compared to a reference calculation

Predicting high/low modulation



Random guess

Metrics

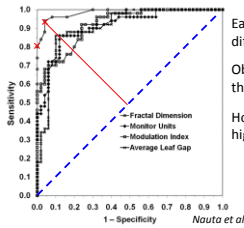
Compare by measuring the area under the curve (AUC)

0.99
0.91
0.91
0.92

Best choice is the fractal dimension

Nauta et al. Med. Phys. 2011

Predicting high/low modulation



Each point on the ROC curve represents a different set of parameters.

Objectively: the best set of parameters maximize the distance from the diagonal

However: a different choice is sometime made (e.g. highest sensitivity for a specificity of 100%)

Nauta et al. Med. Phys. 2011

Importance of a common ground

	McNiven *et al.*	Nauta *et al.*	
[redacted]			✔
Other metrics	Total MU	Total MU, avg. leaf gaps, mod. index	✔
[redacted]			~
Reference	MapCheck result	Fluence smoothing parameter	✖

- A common ground would let us compare studies
  - In this case: how MCS compares to fractal dimension

Some observations

- Designing metrics is easy, designing ***good*** metrics is challenging
  - ROC: easily sorts through potential candidates
- ROC curves can be used to optimize the parameters of a test or classifier
  - However, the range of the parameters must be properly chosen

"it is possible to use the AUC coming from the ROC analysis to determine the best set for these parameters, provided that the range of the parameters is properly chosen."
- Garcia-Romero et al. Med. Phys. 2016

Uses of ROC analysis in the literature

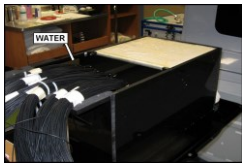
Quantifying detector performance

Quantifying detector performance

- A large diversity of QA detectors exists
  - Is a given detector better at catching some type of errors than other
  - Aside from improved ease of use, is there a point in designing new QA detectors?
- How does a new system compare to older ones?
  - Well demonstrated by the previous presentation

Quantifying detector performance

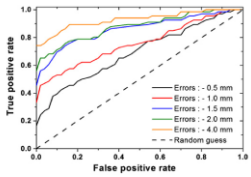
- Example of a new detector:



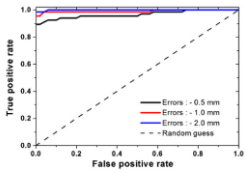
- A plane of 781 scintillating fibers
- Near-perfect water equivalence everywhere
- Currently *a bit* impractical to use

Quantifying detector performance

Single leaf error



Leaf bank error

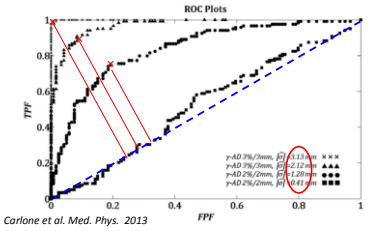


Guillot et al. Med. Phys. 2013

Optimizing detector performance

- For a given detector used for γ evaluation: find the optimal parameters
  - %DD, DTA, threshold
- Example: the MapCHECK
  - Carlone et al. Med. Phys. 2013
  - Sensitivity to leaf errors
    - 17 IMRT plans without error
    - 17 plans with random errors

Optimizing detector performance



Carlone et al. Med. Phys. 2013

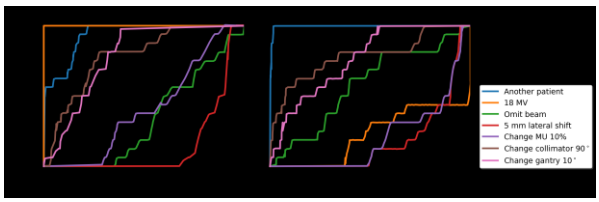
x ideal pass rate for each parameter set

Compromise to be made for different error types

Detection of specific errors

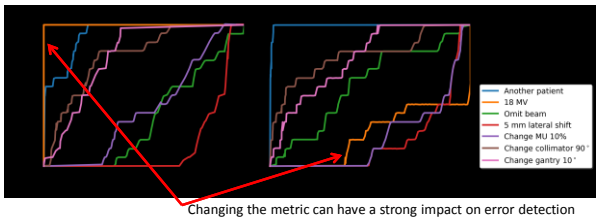
- Childress et al. Med. Phys. 2005
  - Dose calculation with and without errors
    - Wrong energy, wrong patient, collimator/gantry offset, missing beam, MU offset
  - Gamma based analysis
    - 5 % / 3 mm, 3 % / 2 mm
    - Normalized agreement test (NAT), NAT normalized to average PTV dose, γ pass rate …

Detection of specific errors



Data shamelessly extracted from Childress et al., Med. Phys. 2005

Detection of specific errors



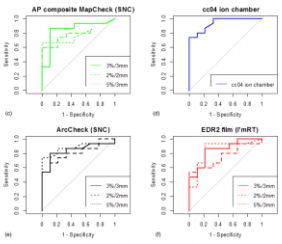
Changing the metric can have a strong impact on error detection

Data shamelessly extracted from Childress et al., Med. Phys. 2005

QA performance assessment

- Other groups have done similar work
  - McKenzie et al. Med. Phys. 2014
    - Extensive study of multiple detectors, %DD/DTA, anatomical site
    - In-house phantom as the reference
  - Sjölin et al. Phys. Medica. 2016
    - Detection incorrect dosimetric leaf gap
  - Bojechko et al. Med. Phys. 2015
    - In vivo EPID
    - MU scaling, MLC noise (random and systematic), patient shift

QA performance assessment



Comparing 'gold standard' in house QA system with various commercial solutions

McKenzie et al. Med. Phys. 2013

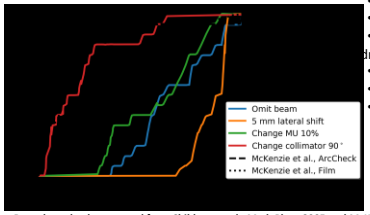
Importance of a common ground (2)

McKenzie et al.

- No specific errors
- Compare measurements
- Gamma pass rate, 5 % / 3 mm

dress et al.

- Specific errors
- Compare dose calculation
- Gamma NAT, 5% / 3 mm



Data shamelessly extracted from Childress et al., Med. Phys. 2005 and McKenzie et al., Med. Phys. 2014

Uses of ROC analysis in the literature

Improving pre-treatment QA

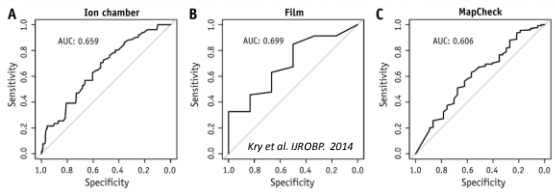
Critique of current pre-tx approaches

- There are numerous critique of gamma based QA:
  - "gamma scores could not reliably identify a plan with poor dosimetric accuracy"
    - Kruse et al. 2010
  - "*planar IMRT QA passing rates do not predict clinically relevant patient dose errors*"
    - Nelms et al. Med. Phys. 2011
  - "*For the same pass-rate criteria, different devices and software combinations exhibit varying levels of agreement*"
    - Hussein et al. Radiother. Oncol. 2013
  - …

Pre-tx QA vs external verification

- Kry et al. (IJROBP 2014) looked at pre-tx QA versus IROC-Houston phantom results
- ROC analysis for 3 type of detectors:
  - **MapCHECK**, **Film**: γ 3 % / 3 mm
  - **Ion chamber**: dose difference
- Does the pre-tx QA predict the phantom results?

Pre-tx QA vs external verification



- Results are slightly better than a random guess
  - QA processes with larger AUC are needed

Toward better and more useful QA

- No matter what the future of pre-tx QA is, it is important to have quantitative assessment of QA systems
  - Large AUC
  - Sensitivity/specificity
    - For different type of errors
  - Optimum parameters

> *"only once [the errors] are detected can they be properly diagnosed and rooted out of the system"*
>
> - Nelms et al. Med. Phys. 2013

Bringing ROC in the clinical routine?

ROC analysis for better QA

- Nobody likes doing useless work
  - Nevertheless, there is increasing evidence that QA may not always provide adequate information
- ROC may address some of these problems
  - Reduce heterogeneity in QA performance
    - Between equipment, institutions
  - Move toward 'evidence based' QA procedures
  - Improve the performance (and usefulness) of QA

How to (as a single institution)?

- An overview of the workflow
  1. Plan list (w/ and w/o errors)
  2. Measurements with a given QA system
  3. Sweep parameters (pass rate, %DD/DTA, …)
     - Classify each plan according to these sets of parameters
  4. Plot ROC curves
  5. Compute AUC
     - Performance of the system VS others
  6. Determine the best set of parameters

How to (as a single institution)?

- While apparently simple, rigorous ROC analysis can be demanding
  - Better results with **lots** of plans with and without errors
- Possible solutions
  - Retrospective analysis
    - But be careful about the reference
  - Scripts/automation to plan and deliver erroneous dose distribution

How to (as a profession)?

- Establish common ground for comparison
  - **What should we use as the reference?**
    - Planned dose distribution
    - Measured plan without errors
    - Results from a given QA system
  - Should we define specific sets of errors to test?
    - If so, which errors?
- Having the same framework will simplify comparison

How to (as a profession)?

- Publish results
  - The more data out there the better
- Provide tools/datasets to ease implementation
  - Script that add errors in DICOM-RT plans
    - Easily done in python
  - Linac automation to run batches of tests
  - Open datasets of plans with and without errors?

Next steps …

- In my opinion, the next steps should be:
  - Try define a common ground
  - Get more papers/data out
  - Make informed decision based on quantitative assessment

… and more distant future

- Pre-tx is essentially a classification task
  - Is the plan good or bad?
- A single passing rate threshold is rather simplistic
  - Machine learning proposes several 'classifiers' that could be trained on our data
  - ROC analysis is the tool of choice to compare classifiers
    - *T. Fawcett, "An introduction to ROC analysis",* ***Pattern Recognition Letters****, 2006*

Last words

- We don't have all the answers yet, but hopefully you are now somewhat convinced of the benefits of ROC analysis to improve our QA
- Lets discuss …