Statistical Considerations When Using MRMC Studies to Evaluate Medical Devices that Incorporate Artificial Intelligence

Changhong Song FDA/CDRH

Outline

- Background
- Study Designs and Statistical Methods
- Examples
- Summary

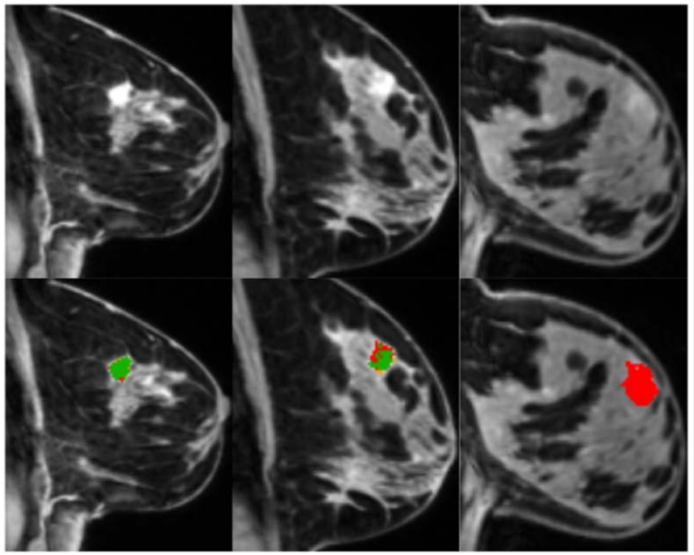
AI / ML

- **Artificial Intelligence** has been broadly defined as the science and engineering of making intelligent machines, especially intelligent computer programs (McCarthy, 2007). Artificial intelligence can use different techniques, including models based on statistical analysis of data, expert systems that primarily rely on if-then statements, and machine learning.
- **Machine Learning** is an artificial intelligence technique that can be used to design and train software algorithms to learn from and act on data. Software developers can use machine learning to create an algorithm that is 'locked' so that its function does not change, or 'adaptive' so its behavior can change over time based on new data.

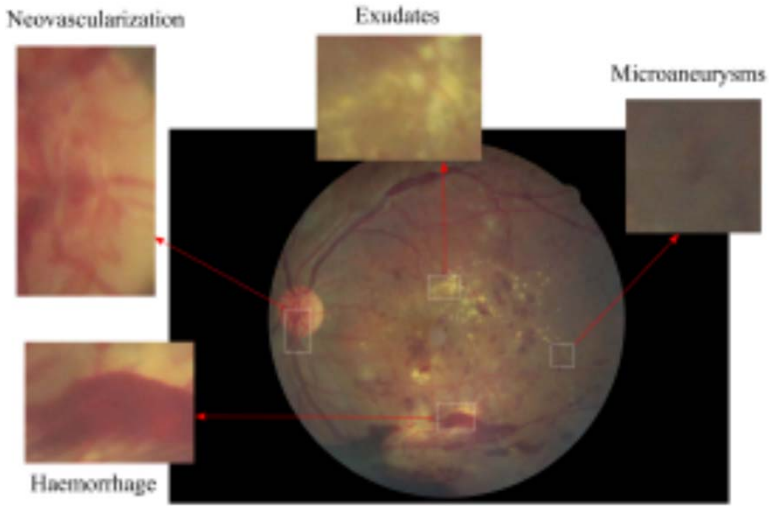
(https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device)

Examples

Computer aided segmentation and diagnosis (Image from Vogl et al. European Radiology Experimental. 3:18. 2019)



Detection of Diabetic Retinopathy on Retina Fundus Images (Image from Islam et al.. 2019. https://arxiv.org/abs/1812.10595)



Multi-Reader Multi-Case (MRMC) study

- Multi-Reader Multi-Case (MRMC) study is a clinical performance assessment study. It has been used for evaluating devices in the areas such as radiological imaging, digital pathology, etc.
- The study design generally involves
  - Multiple readers (intended users): The number of readers needed depends on reader variability.
  - Multiple cases (representative of the patient population)
  - Multiple reading conditions or modalities (e.g., readers unaided versus readers aided by the device)

Multi-Reader Multi-Case (MRMC) study (Continued)

- The study design can be fully-crossed (all readers independently read all cases) or non fully-crossed (e.g. split plot design; Readers read their own group of cases). A fully crossed design has the greatest statistical power.
- Outcome:
  - Binary assessment (diseased or not)
  - An ordinal score (e.g. 0-100%)
  - …

Performance Evaluations

- Standalone Performance:
  - Performance of the device by itself.
- Clinical performance :
  - Clinical performance of the device under intended use by the intended users following device labeling and instructions for use.

Statistical Methods

- Receiver operating characteristic (ROC) curve
- Area Under the ROC Curve (AUC)
- Sensitivity/Specificity
- PPA/NPA
- …

Correlations in the MRMC Study

- Between modalities: the same reader read the same tests
- Between readers: Readers read the same case under the same or different modalities
- Between cases: Cases read by the same readers under the same or different modalities.
- Appropriate statistical methods (e.g. bootstrap methods, model based analysis) should be used to account for the correlated data.

Illustrating Examples

- The study design and statistical analysis will vary depending on device intended use.
- Two examples in different areas will be used to illustrate the statistical methods where MRMC studies are used to evaluate devices with AI algorithm incorporated.

Example 1

Device: A computer-assisted detection (CAD) software device intended to be used concurrently by radiologists while reading digital breast tomosynthesis (DBT) images

Study Design:

- Retrospective study
- # of Readers: 20
- # of Cases (Patients): 240
- two arms (Read with CAD, Read without CAD)
- Fully crossed cross-over design, each reader read every case in each of two arms
- Ground truth: combination of biopsy results and follow-up

Statistical Analysis

- AUC for each reader and their averages over readers were computed with/without CAD using per-subject Level of Suspicion (LOS) score (ranges 0 through 100%).
- Hypothesis for one of the coprimary study endpoints:

H_0: AUC(with CAD) - AUC (without CAD) $\leq -\delta$

H_a: AUC(with CAD) -AUC (without CAD) $> -\delta$

$\delta = 0.05$, pre-specified non-inferiority margin

Statistical Model

$\Theta ijk_{ijk} = \mu + \alpha_i + R_j + C_k$
$+(\alpha R)_{ij} + (RC)_{ik} + (\alpha C)_{jk} + (\alpha RC)_{ijk} + \varepsilon_{ijk}$

Θijk: an estimate of AUC

μ: the population mean

α : fixed effect of modality i (i=1 with CAD, 2 without CAD)

R: random effect of readers ~ $(0, \sigma_R^2)$

C: random effect of cases $\sim (0, \sigma_C^2)$

αR, αC, RC, αRC: interaction terms with mean 0 and variances of $\sigma^2_{\alpha \mathrm{R}}$, $\sigma^2_{\alpha \mathrm{C}}$, σ^2_{RC} and $\sigma^2_{\alpha \mathrm{BC}}$

ε_{ijk}: random error $\sim (0, \sigma_\varepsilon^2)$

Results

Average AUC (with CAD)	**0.850**
Average AUC (Without CAD)	0.841
AUC Difference	0.009
95% CI	(-0.012, 0.030)
P-value	< 0.01

https://www.accessdata.fda.gov/cdrh_docs/pdf16/P160009B.pdf

FDA Guidances about Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device

- "Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions - Guidance for Industry and Food and Drug Administration Staff"
- "Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions - Guidance for Industry and FDA Staff"

Example 2 (Hypothetical)

Device: A device with AI algorithm that will aid Pathologists in identifying cancer in digitized biopsy slides by selecting areas in the slide that are suspicious.

Study Design:

- Multiple Readers (20)
- Multiple Cases: (600; 300 positives/300 negatives)
- Arms/ Modalities (manually review using the glass slides (MR)/readers read digital slides directly (DR), readers read digital slides with AI recommendation(Readers + AI), AI prediction by itself (AI))
- Fully crossed study design (each reader read every case) under all modalities
- Case will be classed as either cancer or no cancer by the device and pathologists.

AI Algorithm Evaluation

Clinical (Reader+ AI) Performance

- Pathologists + AI v.s. Manual read (MR) by pathologists or
- Pathologists + AI v.s. Digital read (DR) by pathologists
- …

Standalone Performance

- AI Prediction with a performance goal
- AI Prediction v.s. Manual read (MR) by pathologists or
- AI Prediction v.s. Digital read (DR) by pathologists
- …

Study Data

- For the modalities with reader interpretation, the data will be collected for each case and reader for each modality (e.g. 1 denotes cancer and 0 denotes non-cancer).

	Reader 1	…	…	Reader n
Case 1	1	1	…	0
…	0	0		0
…	1	1	1	1
…				
Case m	1	1	1	1

- The modality with AI prediction will have one prediction for each case.

With Reference Standard

Sensitivity (Se)= Prob (test=+ | reference=+)

=True Positive/(True Positive + False Negative)

Specificity (Sp)= Prob (test=- | reference=-)

= True Negative/(True Negative + False Positive)

Test whether sensitivity/specificity of one modality is superior or non-inferior to the other modality or test whether sensitivity/specificity can meet clinically acceptable performance goal.

E.g.

Modality 1 = Comparator, Se1 & Sp1

Modality 2= New device , Se2 & Sp2

Superiority on sensitivity and non-inferiority on specificity:

H_0: $(Se_2 - Se_1) \leq \delta 1$ $\quad H_a$: $(Se_2 - Se_1) > \delta 1$

H_0: $(Sp_2 - Sp_1) \leq -\delta 2$ $\quad H_a$: $(Sp_2 - Sp_1) > -\delta 2$

Where $\delta 1 \geq 0, \delta 2 > 0.$

Without Reference Standard

- Choice of Comparators, for examples
  - Manual read (MR)
  - Cleared/approved Devices

- PPA and NPA can be used to summarize agreement of the new device results with comparator results.
  - PPA = Prob (device = + | comparator=+)
  - NPA = Prob (device = - | comparator=-)
  - Study Hypothesis :

$$H_0: PPA \leq P_0 \qquad H_a: PPA > P_0$$
$$H_0: NPA \leq P_1 \qquad H_a: NPA > P_1$$

P_0 and P_1 are generally pre-specified based on clinical acceptability.

Model based analysis

Model based on analysis can be performed to account for correlations.

$$\text{logit}(p_{ijk}) = \mu + \alpha_i + R_j + C_k + (\alpha R)_{ij} + (RC)_{ik} + (\alpha C)_{jk} + (\alpha RC)_{ijk}$$

p_{ijk}: proportion of agreement with reference standard (sensitivity, specificity)
μ: the population mean
α: modality, i=1, 2, 1 for comparator, 2 for new device
R: random effect of readers ~ (0,σ_R^2)
C: random effect of cases ~ (0,σ_C^2)
αR,αC, RC , α$R$$C$: interaction terms with mean 0 and different variances of $\sigma^2_{\alpha R}$, $\sigma^2_{\alpha C}$, σ^2_{RC} and $\sigma^2_{\alpha BC}$

Sensitivity: use all the positive cases by reference standard for analysis,
Specificity: use all the negative cases by reference standard for analysis

Summary

- AI / ML related medical devices are getting more and more applications.
- MRMC studies have been used to evaluate devices in the areas such as radiology and digital pathology where reader interpretation of the imaging results is needed.
- Depending on device intended use, the study design and statistical analysis methods can be quite different. The design of the MRMC study can be affected by the outcome types of the device and clinical interpretations.

Acknowledgement

- Dr. Xiaoqin Xiong
- Dr. Yaji Xu
- DCEA2: Division of Biostatics at FDA/CDRH

Related FDA Guidances

- “Design Considerations for Pivotal Clinical Investigations for Medical Devices” FDA Guidance
- “Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests”
- “Software as a Medical Device (SAMD): Clinical Evaluation”
- “Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions - Guidance for Industry and Food and Drug Administration Staff”
- “Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions - Guidance for Industry and FDA Staff”

  - https://www.fda.gov/regulatoryinformation/guidances/