FDA Perspectives on Machine Learning / Deep Learning Devices for Medical Image Interpretation

Berkman Sahiner, PhD
FDA/CDRH/OSEL/DIDSR
Outline

- FDA premarket process for medical devices
- Machine learning for image interpretation
  - FDA guidances on CADe
- Some common pitfalls in device submissions involving machine learning
- Adaptive systems
- DIDSR research related to machine learning
FDA Premarket Process for Medical Devices
What is a Medical Device?

• Regulated under authority of Federal Food, Drug, and Cosmetic Act (Section 201h)
  – “… an instrument, apparatus, implement, machine, … intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment or prevention of disease, in man or other animals, or intended to affect the structure or any function of the body of man or other animals …”
Medical Device Classification

• Intended Use / Indications for Use
• Risk-Based Paradigm
  – Medical devices are classified and regulated according to their degree of risk to the public
## Device Class and Pre-Market Requirements

<table>
<thead>
<tr>
<th>Device Class</th>
<th>Controls</th>
<th>Premarket Review Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I (lowest risk)</td>
<td>General Controls</td>
<td>Most are exempt</td>
</tr>
<tr>
<td>Class II</td>
<td>General Controls &amp; Special Controls</td>
<td>Mostly Premarket Notification [510(k)]</td>
</tr>
<tr>
<td>Class III (highest risk)</td>
<td>General Controls &amp; Premarket Approval</td>
<td>Mostly Premarket Approval [PMA]</td>
</tr>
</tbody>
</table>
General Controls

- All or some may be required, depending on device class
  - Establishment registration and device listing
  - Good Manufacturing Practices (GMPs)
  - Labeling
  - Record keeping
  - Reporting of device failures
  - Prohibition against adulterated or misbranded devices
  - Premarket notification 510(k)
Special Controls

• Regulatory requirements for class II devices
  – Devices for which general controls alone are insufficient to provide reasonable assurance of the safety and effectiveness of the device

• Generally device-specific

• Examples of special controls
  – Premarket data requirements
  – Performance standards
  – Guidelines
  – Special labeling requirements
  – Postmarket surveillance
510(k) Premarket Notification

• Path to market for the majority of medical devices
• Requires determination that a new device is substantially equivalent to a legally marketed device (predicate device)
Substantial Equivalence

• Option 1
  – Has the same intended use as the predicate; and
  – Has the same technological characteristics as the predicate;
Substantial Equivalence

• Option 2
  – Has the same intended use as the predicate;
    and
  – Has different technological characteristics and the information submitted to FDA, including appropriate clinical or scientific data where necessary, demonstrates that the device:
    • Does not raise different questions of safety and effectiveness than the predicate; and
    • Appropriate clinical or scientific data demonstrate that the device is at least as safe and effective as the predicate
510(k) Process

- Key content of application
  - Indications for Use Statement
  - Executive Summary
  - Device Description
  - Substantial Equivalence Discussion
  - Proposed Labeling
  - Sterilization and Shelf Life, Biocompatibility, Software, Electromagnetic Compatibility and Electrical Safety
  - Performance Testing – Bench, Animal, Clinical
Performance Data

• Performance data needed to support a premarket submission depends on the intended use of the device

“as a second reader, this device assists radiologists in minimizing oversights by identifying image areas that may need as second review” ≠ “this device analyzes screening mammograms and identifies mammograms that do not need any further review by a radiologist”
**Premarket Approval (PMA)**

- **Class III devices**

- **Demonstrate reasonable assurance of safety and effectiveness**
  - Very device specific
  - Standalone submission
    - No comparison to a predicate

- **When FDA finds a device safe and effective for its intended use, it is “approved”**
PMA Elements

- Summary
- Device Description
- Reference to any standards
- Non-clinical studies
- Clinical investigations involving human subjects
- Proposed Labeling
- Review of quality systems
- Inspection of facility
- Bioreserarch monitoring audit of clinical data sites
De Novo

- Novel devices that have not previously been classified are Class III by default (and hence, PMA devices)
- De novo is a petition for down classification (Class III to Class II or Class I)
- De novo petition must propose controls that would be needed to assure the safety and effectiveness of the device
De Novo

• A granted de novo establishes a new device type, a new device classification, a new regulation, and necessary general (and special) controls

• Once the de novo is granted, the device is eligible to serve as a predicate
  – All the followers are 510(k) devices
PreSubmissions (Qsubs)

- Informal interaction with FDA (usually non-binding)
  - Informational Meeting
  - Study Risk Determination (FDA decision is final)
  - Early Collaboration Meeting
  - Submission Issue Meeting

- Helps avoid delays in device submission or repeating clinical studies
Example Pre Submission Questions

• Does FDA agree that the proposed test plan is sufficient to support the stated indication?

• Is FDA aware of standardized test methods to address concern XYZ?

• FDA has requested that we do X; would the proposed combination of Y and Z be acceptable to the Agency?
More Information

- Please go to [https://goo.gl/eg9hRv](https://goo.gl/eg9hRv) for recent presentations by members of Division of Radiological Health (DRH) and Division of Imaging, Diagnostics and Software Reliability (DIDSR)
Machine Learning for Medical Image Interpretation
Machine Learning in Image Interpretation

- Computer-aided detection (CADe)
  - First reader
  - Sequential reading
  - Concurrent reading
- Computerized detection
- Image interpretation devices used in intra-operative procedures

- Computer-aided diagnosis (CADx)
  - Presence/absence of disease
  - Severity, stage, prognosis, response to therapy
  - Recommendation for intervention
- Quantitative imaging
- Many other possibilities
Computer-Aided Detection

• First device approved in 1998
  – Mammographic CADe

• Additional application areas
  – Chest x-ray
  – Lung CT
  – Colon CTC
  – Other areas
CADe Guidances

- Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data – Premarket Notification [510(k)] Submissions
  
  http://www.fda.gov/RegulatoryInformation/Guidances/ucm187249.htm

- 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
  
  http://www.fda.gov/RegulatoryInformation/Guidances/ucm187277.htm
Content of 510(k) Guidance

- General device information
- Technological Characteristics
  - Algorithm design and function
  - Processing steps
  - Features
  - Models and classifiers
  - Training paradigm
- Databases
- Reference standard
- Scoring methodology
- Assessment
Device Description - General Information

- Target population
- Type of data compatible with CAD design
- Workflow
- Device impact: Both TPs and FPs
- Device limitations
- Supporting literature
Technological Characteristics

• Needed for
  – Understanding scope of a change in a modified CADe algorithm
  – Comparing two devices
  – May reduce performance testing requirements
Algorithm Design and Function

• Flowchart describing processing, features, models, classifiers

• How are algorithm parameters selected?

• Better description increases confidence in the device
Databases

• May contain computer simulated data, phantom data, or patient data

• Phantom data
  – Phantom or simulation methodology
  – Data characterizing the relationship between the simulated or phantom data and actual patient data
Databases

• Patient data
  – Demographics
  – Conditions of radiologic testing
  – How the imaging data were collected
  – Collection sites
  – Number of cases, stratification
  – Comparison of characteristics of the patient data to the target population
  – History of accrual and use of both training and test
Reference Standard

• How is disease presence/absence determined for the cases in the sponsor’s database?
  – Output from another device
  – An established clinical determination (e.g., biopsy)
  – A follow-up examination
  – An interpretation by reviewing clinicians (truthers)
    • number and qualifications of truthers
    • specific criteria used as part of truthing
    • other clinical information utilized
Scoring Methodology

• Correspondence between CADe output and reference standard

• Example
  – Device TP
    • CAD mark within a certain distance from the reference standard
  – Device FP
    • CAD mark away from the reference standard

• Scientific justification
Assessment

• Standalone assessment:
  – Report the performance of the CADe device by itself, in the absence of any interaction with a clinician

• Clinical Performance Assessment:
  – Demonstrate the clinical safety and effectiveness of the device for its intended use, when used by the intended user and in accordance with its proposed labeling and instructions
  – May not be necessary if direct comparison to a legally marketed predicate device is available
  • Same evaluation process and same test data set.
Standalone Assessment

• Dataset can be larger than clinical assessment
  – Evaluate performance on different subsets
    • Acquisition parameters
    • Lesion properties
      – Size, severity, etc.
    • Powering for subsets is not necessary unless specific claims

• Training / Test set hygiene
  – Generalizability
  – Danger of overtraining
Standalone Assessment – Test Data Reuse

• Permitted with tight controls:
  – Data randomly selected from a larger database that grows over time
  – Fixed limit on the number of times a case can be used for testing
  – Tight control on data access:
    • Algorithm development team does not have access to data
    • Only summary performance results are reported outside of the assessment team
  – Data access log is maintained
Standalone Testing - Detection Accuracy

• Sensitivity AND

• FPs or other measure of specificity
  – Confidence intervals

• FROC (Free-response ROC) analysis
  – Confidence intervals
Standalone Testing - Comparison to Predicate

- Difference in area under the FROC curves OR
- Difference in sensitivity and FPs/case
Clinical Assessment
510(k), PMA, or De Novo

• Why needed:

  – Reader is an integral part of the diagnostic process for CADe devices
  – Comparison of overall standalone performance may not be adequate when
    • Substantive technological differences with predicate, resulting in marks with substantially different characteristics
    • Differences in indications for use
Clinical Study Design

• Prospective study
  – May not be practical

• Retrospective MRMC study
  – Multi-reader multi-case
MRMC Study Protocol

• Hypothesis to be tested

• Statistical and clinical justification of
  – Case sample size
  – Number of readers

• Image interpretation methodology and relationship to clinical practice

• Randomization methods

• Reader task
Statistical Analysis Plan

• Pre-specification is a key factor

• Some elements (list not exhaustive)
  – Endpoints, statistical analysis methods
  – Process for defining truth (reference standard)
  – Scoring technique
  – Any results from a pilot study supporting the proposed design
  – Plan for multiple hypothesis testing if appropriate
  – Plan for handling missing data
Subset Analysis

• Sufficient number of cases from important subsets to estimate standalone performance and confidence intervals

• Powering subsets for statistical significance not necessary unless specific subset performance claims are included
Some Common Pitfalls
Training / Test Dataset Hygiene

• Test dataset characteristics may leak into classifier design
  – Often inadvertently

• Simple example:
  – Design $N$ classifiers with different parameters on training set
  – Apply each classifier to the test set
  – Submit the best result as the test result to FDA
  – Don’t do it!
    • (without an appropriate statistical plan)
Image Acquisition Device / Parameter Range too Wide OR Too Narrow in Device Testing

• Range too narrow
  – Not all possible devices / acquisition parameters tested
    • Implications for device labeling

• Range too wide
  – Possibility to fail to achieve desired endpoint

• Can’t I first look at my final pivotal study results and then decide the range that I will specify in my device labeling?
  – Remember that pre-specification of analysis plan is required
Multiple Hypothesis Testing without an Appropriate Statistical Plan

• Example:
  – H1: My device improves the radiologists’ performance in disease detection
  – H2: My device increases radiologists’ reading time with comparable performance to not using the device

• What happens if I
  – Test H1 and H2 in a single experiment
  – Declare a significant test outcome if one of the two hypothesis is “locally” significant?
    • $p_1<0.05$ or $p_2<0.05$
Multiple Hypotheses

• Assume in reality the device is not useful for either purpose

• What is the probability of falsely declaring success?
  – $p(\text{reject at least one null hypothesis})$
    
    $=1-p(\text{fail to reject both null hypotheses})$
    
    $=1-0.95 \times 0.95 = 0.0975$ (if $H_1$ and $H_2$ are statistically independent)

• “Type-I error rate” is inflated

• Statistical plan expected to prospectively account for multiple hypothesis testing
Testing Does not Match Intended Use

• A device that claims to improve radiologists’ performance
  – Testing includes only standalone results
  – There is no scientific evidence that a certain standalone performance level will improve the user’s performance
Test Population Does not Match Intended Use Population

- Using the submitted device to select the test population
  - Don’t do it!

- Patient population is not representative
  - Some of the relevant lesions/patients that meet inclusion criteria were later excluded
  - Need prospective plan about how to handle missing data

- Dataset enrichment and stress testing
  - Discuss in QSub
Adaptive Systems
Adaptive Devices

• Device algorithm evolves over time based on the new evidence collected in the field after the device goes on the market
Current Guiding Principles for Changes in 510(k) Devices

• If a manufacturer modifies their device with the intent to significantly improve the safety or effectiveness of the device a new 510(k) is likely required

• Software change guidance
  – Need new 510(k) if
    • Software change could significantly affect clinical functionality or performance specifications directly associated with intended use
Adaptive versus Traditional Changes

**Traditional**

1. Data Collection
2. Training Set
3. Algorithm Development
4. Device Algorithm v1
5. Performance Testing
6. Enlarged Training Set
7. Algorithm Development
8. Device Algorithm v2
9. Performance Testing

A few months after device on the market
Adaptive versus Traditional Changes

Traditional

Data Collection → Training Set → Algorithm Development → Device Algorithm v1 → Performance Testing

A few months after device on the market → Enlarged Training Set → Algorithm Modification → Device Algorithm v2 → Performance Testing

Adaptive

Data Collection → Training Set → Algorithm Modification → Device Algorithm v1 → Performance Testing

A few days after device on the market … → Algorithm Modification → Device Algorithm v2 → Performance Testing
Questions Regarding Adaptive Devices

• How can we ensure that the changes do not compromise the safety and effectiveness of the device?
  – Risk associated with the device
  – Extent of automated adaptation

• Can and should performance testing be conducted after each algorithm modification?

• Test dataset reuse
Training to the Test

• If you get to interrogate the test set many times, your so-called “test” set becomes part of training
  – Performance seems to improve for the so-called “test” set as the device adapts
  – If you collect a fresh sample from the true population, you find that performance has actually declined
Freedman’s Paradox

Design #1

Training Set → Feature Selection → Classifier Design → Finalized Classifier → Performance Testing → Holdout Set

Report Device Performance
Freedman’s Paradox

Design #2

- Training Set
- Feature Selection
- Classifier Design
- Finalized Classifier
- Performance Testing
- Holdout Set
- Report Device Performance
Design #2

Freedman’s Paradox

- Training Set
- Selected Features
- Classifier Design
- Finalized Classifier
- Performance Testing
  
  Reported Device Performance
  
  Holdout Set
Training to the Test: Useless Variables

Number of Variables

Accuracy

0.50 0.55 0.60 0.65 0.70

0 200 400 600 800 1000

Training Holdout Fresh

Standard Holdout

Adapted from Dwork et al. Science 2015;349:636-638
Training to the Test

• There may be ways to mitigate this
  – At the expense of increased uncertainty
  – CAD Guidance: Data randomly selected from a larger database that grows over time

• Dwork et al., Science 2015
  – In essence, add noise to reported performance each time you access the test set
  – Differential privacy from cryptography
Useless Variables

Thresholdout

Accuracy

Number of Variables

Training
Holdout
Fresh

Adapted from Dwork et al. Science 2015;349:636-638
DIDSR Research Related to Machine Learning
Division of Imaging, Diagnostics and Software Reliability

- Part of the Office of Science and Engineering Labs (OSEL)
- Support the mission of OSEL through investigating issues related to
  - Medical imaging
  - Computer-assisted diagnostics
  - Software reliability
OSEL Mission

• Conduct laboratory-based regulatory research to facilitate the development and innovation of safe and effective medical devices

• Provide expertise, data, and analyses to support regulatory process

• Collaborate with academia, industry, government, and standards development organizations to develop, translate, and disseminate information regarding regulated products
DIDSR Github Site

- Make a range of simulation and analysis tools, and data available to the public

https://github.com/DIDSR
Some Research Topics

• Lesion insertion for dataset augmentation
• Calibration
• Role of the precision-recall curve for assessment
Lesion Insertion for Dataset Augmentation

• Lack of large, annotated medical image data sets with lesions
  – Major bottleneck in
    • Development
      – Think deep learning
    • Assessment of CAD systems

• Design and evaluate computational techniques to insert realistic lesions into clinical images
  – Lesion location perfectly known
  – Properties known about the “original” lesion typically transferred to the inserted lesion
    • Size
    • Malignant / benign
    • Pertinent lesion characteristics
Lesion Insertion

- Seamlessly insert real, imaged lesions into normal images

Source Image
Pezeshk et al., IEEE Trans. Biomedical Eng., 2015

Target Image

Direct Insertion

Insertion w/o special processing

Final Method
Reader Studies for Algorithm Validation

- Task: Provide score for realism
  - 100: Definitely real
  - 0: Definitely inserted

- Figure of merit
  - Area under the ROC curve (AUC)
  - AUC=0.5: Insertion is perfect
  - AUC=1.0: Insertion is awful

<table>
<thead>
<tr>
<th>Mammography Data Set</th>
<th>Years of Experience</th>
<th>AUC ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>17</td>
<td>0.53 ± 0.06</td>
</tr>
<tr>
<td>R2</td>
<td>30</td>
<td>0.56 ± 0.07</td>
</tr>
<tr>
<td>Average</td>
<td>23.5</td>
<td>0.54 ± 0.05</td>
</tr>
</tbody>
</table>

- Radiologists have a hard time distinguishing real and inserted lesions

Pezeshk et al., SPIE MI 2016, 97850J
Application to CAD Training

• TR1: Train with original lesions
• TR3: Train with augmented data
  – Augmentation includes transformation and insertion

Now we are ready to apply the lesion insertion methods to device evaluation
  – Mammography CAD
  – Estimation of performance of CAD with a new mammography acquisition device
  – Reduce clinical data acquisition needs with new device
Calibration

• Many algorithms provide scores on an arbitrary scale
  – Difficult to interpret
  – Especially when there are two or more devices
• For a calibrated device:
  • A device output of x\% means that
    – x of 100 patients actually have the outcome

• Three calibration methods
  – Uncertainty estimation and
  – Simulation studies

• Under rationality assumption, calibration is independent of discrimination
Precision-Recall Curve

• Alternative performance evaluation technique
  – Information retrieval systems
  – Content-based image retrieval systems

• AUCPR: Area under the precision-recall curve

• Limited work in investigating variance

Giger et al, SPIE Medical Imaging 2002 4684:768-773
AUCPR

- Proposed a new method for estimating AUCPR
  - Inherits many nice features of semiparametric estimation of the area under the ROC curve
  - Can be used for system assessment and statistical comparison

Sahiner et al., SPIE Medical Imaging 2016 97870D
Summary

• Performance data needed to support a premarket submission depends on the submission type, the device technology, and the intended use of the device

• Device manufacturers are encouraged to use the pre-submission mechanism to seek FDA input

• Adaptive systems pose new challenges

• FDA researchers are active in issues related to machine learning and we make a range of tools, and data available to the public
Acknowledgements

• Special thanks to many members of CDRH, and in particular to
  – Weijie Chen PhD
  – Jana Delfino PhD
  – Brandon Gallas PhD
  – Aria Pezeshk PhD
  – Nicholas Petrick PhD

for their input to this presentation