Pre- and Post-Market Evaluation of Autonomous AI/ML: Lessons Learned from Prior CAD Devices

Nicholas Petrick, Ph.D.

Division of Imaging, Diagnostics and Software Reliability (DIDSR)
Office of Science and Engineering Labs (OSEL)
Center for Devices and Radiological Health (CDRH)
U.S. Food and Drug Administration (FDA)
nicholas.petrick@fda.hhs.gov

OUTLINE

- CAD devices types
- CAD assessment framework
- Some lessons learned
  - and how they might apply to autonomous AI/ML

COMPUTER-AIDE DIAGNOSIS (CAD)

CADe	• Computer aided detection: Aids in localizing/marking regions of that may reveal specific abnormalities
CADx	• Computer aided diagnosis: Aids in characterizing/assessing disease, disease type, severity, stage, progression
CADe/x	• Computer aided detection/diagnosis : Aid in localizing and characterizing conditions
CADt	• Computer aided triage: Aids in prioritizing/triaging time sensitive patient detection and diagnosis
CADa/o	• Computer-aided acquisition/optimization: Aid in the acquisition/optimization of images/diagnostic signals

These CAD device types are general not organ/disease specific

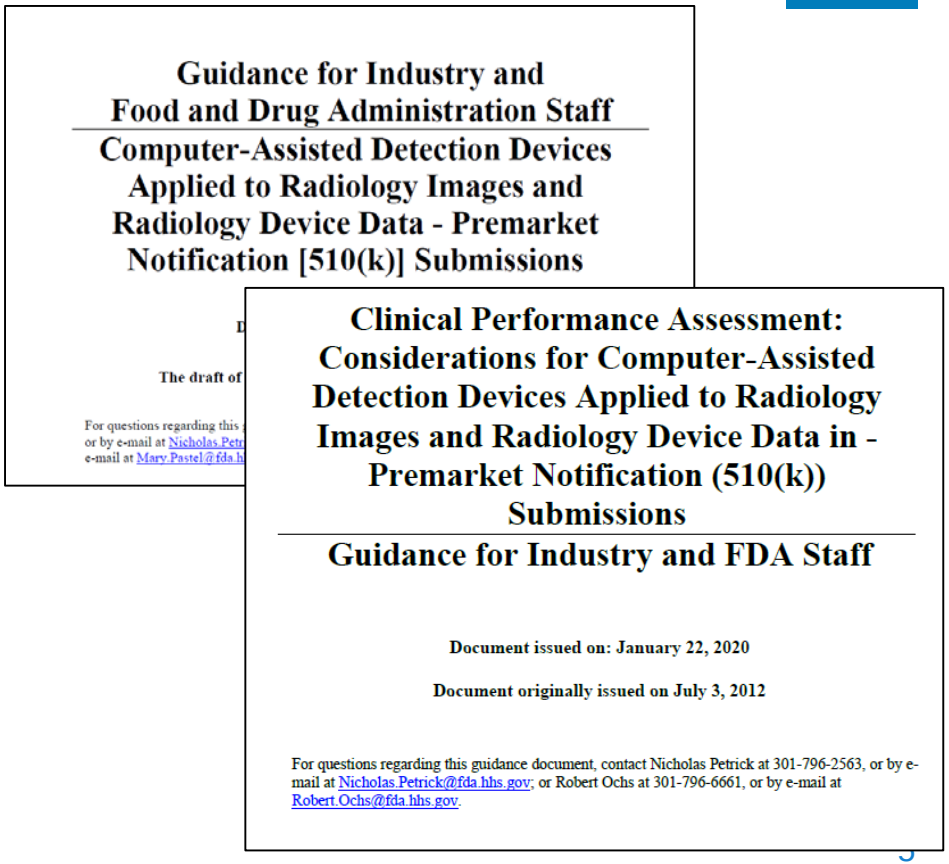
Initial Screener

- Initial Screener
  - Automated cervical cytology slide reader
    - BD FocalPoint GS (P950002-S002)
      - Selects up to 25% of Pap smear slides that need no further review

*https://www.bd.com/en-us/offerings/capabilities/cervical-cancer-screening/cytology-instruments/focalpoint-gs-imaging-system

CAD Device Assessment

- Two CADe guidance documents
  - Clinical assessment guidance
    - Updated 1/22/2020 to reflect CADe reclassifications



Guidance for Industry and
Food and Drug Administration Staff

Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions

The draft of

For questions regarding this
or by e-mail at Nicholas.Petr
e-mail at Mary.Pastel@fda.h

Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in - Premarket Notification (510(k)) Submissions

Guidance for Industry and FDA Staff

Document issued on: January 22, 2020

Document originally issued on July 3, 2012

For questions regarding this guidance document, contact Nicholas Petrick at 301-796-2563, or by e-mail at Nicholas.Petrick@fda.hhs.gov; or Robert Ochs at 301-796-6661, or by e-mail at Robert.Ochs@fda.hhs.gov.

Basic CADe Assessment Framework

- Detailed Descriptions
- Verification and Validations (V&V)
  - Standalone performance assessment
  - Clinical performance assessment
  - …
- Labeling
- …

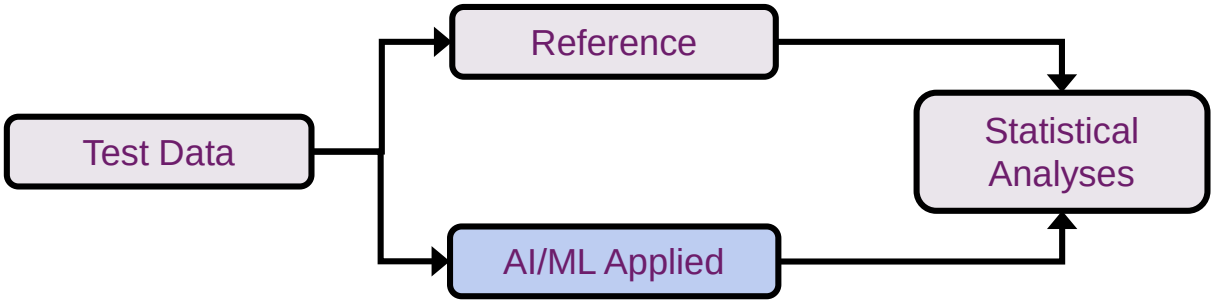
*Based on information in the CADe Guidances

Detailed Descriptions

- CAD algorithm and component
  - Device usage and patient population
  - Algorithm design and function
  - Training process
    - Training/tuning datasets
  - …
- Performance testing protocols and dataset(s)
  - Statistical analysis plan
  - Validation dataset(s)
    - Independent of training/tuning datasets & only accessed after algorithm finalized
  - Reference standard
  - Scoring method
  - …

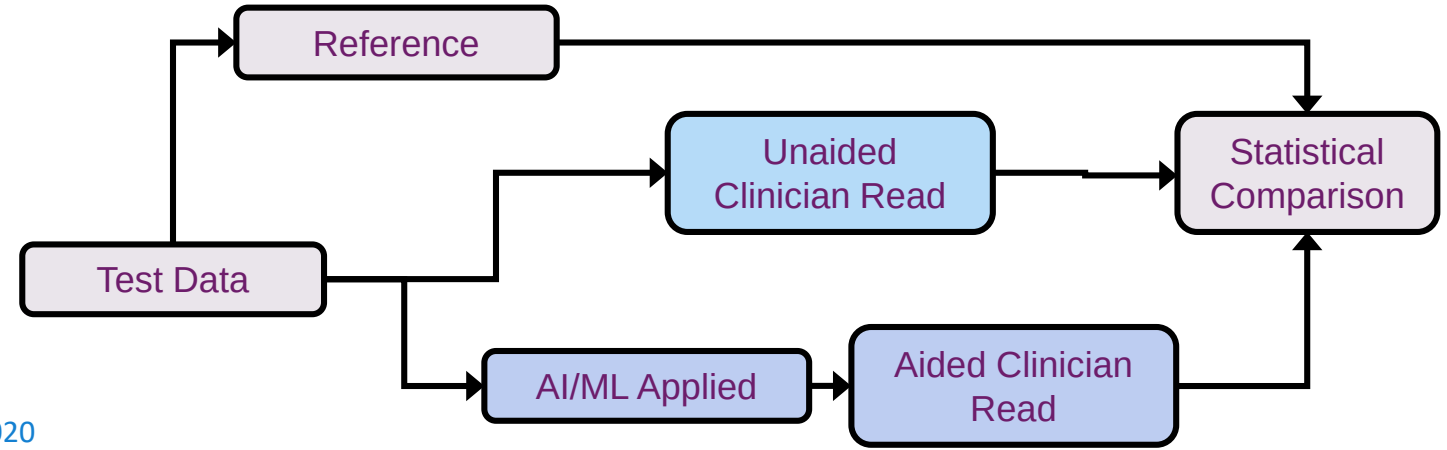
Standalone Performance Assessment

- Benchmark performance (includes CIs)
- Facilitate generalizability analysis



Clinical Performance Assessment

- Clinicians' performance utilizing the device
  - Multi-reader multi-case designs
  - Prospective/retrospective



Other CAD Assessment Frameworks

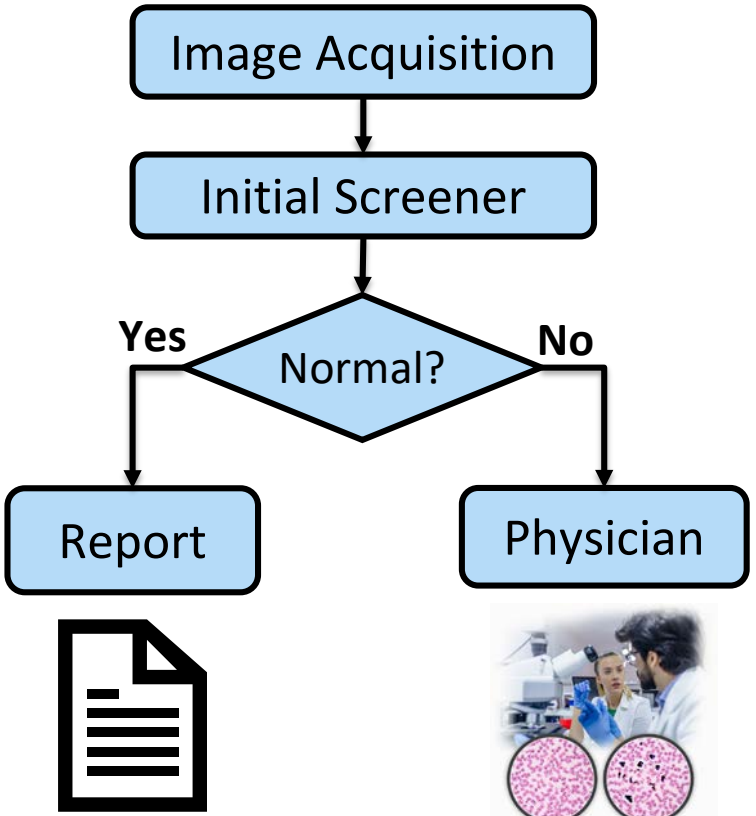
- Special Controls
  - CADx: DEN170022 Quantitative Insights QuantX
  - CADe+CADx: DEN180005 Imagen OsteoDetect
    - Generally consistent with CADe Guidance approaches
  - CADt: DEN170073 Viz.AI ContraCT
    - Detailed Descriptions
      - …
    - V&V
      - Standalone performance testing
      - Performance testing that demonstrate device will provide effective triage
      - …
    - Labeling
  - CADa/o: DEN190040 Bay Labs Caption Guidance

LESSON 1

- CAD assessment framework largely applicable to autonomous AI/ML
  - Important differences
    - CADe/CADx
      - Includes standalone and clinical reader performance
      - Generally utilizes enriched dataset
    - Autonomous AI/ML
      - Include standalone/benchmark performance
      - Include clinical comparison with standard of care (?)
      - Representative dataset likely required

Initial Screener

- Initial Screener (AI/ML+Reader)
  - Normal: AI/ML or reader determine “Neg”
  - Positive: Both AI/ML & reader determine “Pos”
- Properties
  - Se AI/ML+Reader ≤ Se SOC
    - AI/ML may miss cancers
  - Sp AI/ML+Reader ≥ Sp SOC
    - SOC reader may misdiagnose normal cases determined correctly by AI/ML
- Assuming reader performance same before/after introducing initial screener



LESSON 2

- Makeup of validation dataset impacts safety and effectiveness assessment

Common Test Dataset Types

- Enriched datasets
  - Challenging cases to probe differences in modalities
  - Enriched subgroups to assess generalizability
    - Technological subgroups
    - Patient subgroups
- Representative datasets
  - Representative of clinical population
  - Representative of imaging acquisition landscape

CAD Test Datasets

- Evaluating CADs in FDA submissions
  - Generally not measuring aid under clinical conditions
  - Generally not comparing devices across different datasets

  - Standalone
    - Representative (somewhat) to benchmark performance
      - Limited number of sites, likely some selection bias
    - Enriched subgroups/imaging technologies
      - Informs labeling and facilitates generalizability analyses

  - Clinical reader study
    - Typically enriched with challenging cases
    - Clinical readers generally not truly representative of reader population

AUTONOMOUS AI/ML TEST DATASETS

- Evaluating autonomous AI/ML in FDA submissions
  - Likely want ability to direct compare algorithms across different datasets
  - Standalone
    - Representative of patient population and clinical image acquisitions
      - More sites, minimize potential selection bias
    - Enriched patient or imaging technologies subgroups
      - Inform labeling, determine limitations, generalizability analysis
  - Standard-of-Care comparison study
    - Representative of clinical population, clinical image acquisitions, clinical readers
      - Automated Pap smear reader device used a two-arm prospective intended use study at five cytology laboratories

DATASET SIZE

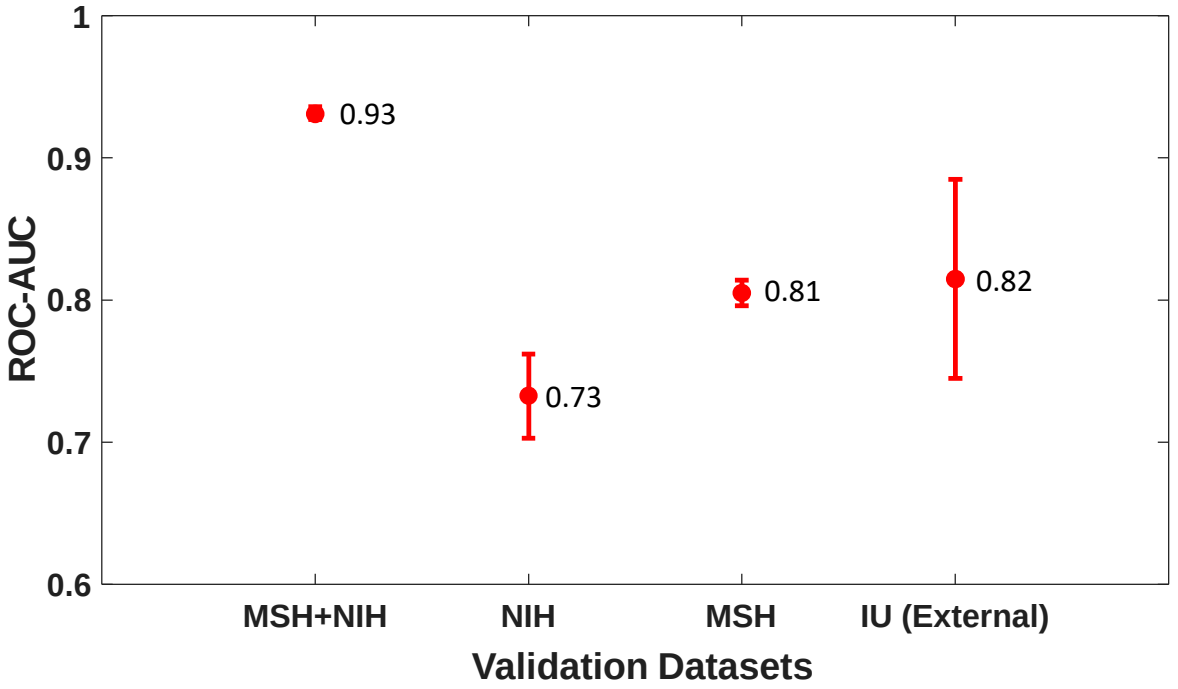
- How much data is enough?

CAD has generally uses modest sized enriched validation datasets

Automatous AI/ML likely to require larger representative validation datasets

LIMITS OF AUTONOMOUS AI/ML TESTING

- Zech et al*, CNN pneumonia screener for chest X-rays
  - Training MSH & NIH
    - Train:109,922 cases
    - Tune: 15,144
  - Validation
    - NIH: 22,062
    - MSH: 8,388
    - IU: 3,807
  - Pneumonia prevalence
    - NIH: 1.2%
    - MSH: 34.2%
    - IU: 1.0%

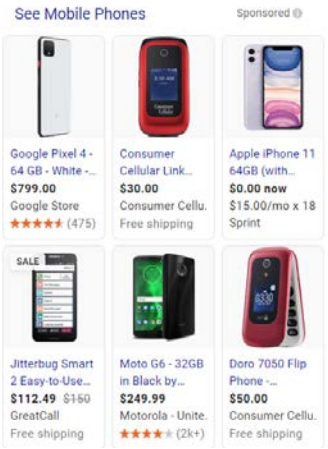


*Zech et al, PLOS-Medicine, 2018

LIMITS OF AUTONOMOUS AI/ML TESTING

- Limits of benchmarking AI/ML performance

Extend to new imaging devices without new studies?



Extend to other patient groups without new studies?

Adults

Children

LESSON 3

- Quality control (QC) is critical for automated AI/ML
  - QC as safety control

AUTOMATED PAP SMEAR INITIAL SCREENER

- Automated QC checks on inputs & AI/ML processing
  - To ensure real-time adequacy of algorithm processing
- Clinical QC
  - To confirm ongoing AI/ML functionality
    - System identifies up to 15% of successfully processed slides for full manual rescreening

Summary of CAD Lessons

- For autonomous AI/ML
  - CAD assessment framework largely applicable
  - Large representative validation dataset likely needed
  - Ongoing quality control may be a useful safety control


FDA
U.S. FOOD & DRUG
ADMINISTRATION