Validation of AI Algorithms in Guided Imaging Applications

Berkman Sahiner, PhD
Division of Imaging, Diagnostics, and Software Reliability
OSEL/CDRH/FDA

Performance Testing:
Non-Clinical and Clinical

- Non-clinical
  - Characterize the technical performance of the AI system for guided image acquisition
- Clinical
  - Evaluate diagnostic utility of the device when representative users use the device on a representative patient population

Why Do We Need Both?

- Non-clinical
  - Can be performed on larger data sets
    - May be able answer questions that require good precision
  - May be helpful for future iterations of the device
- Clinical
  - Quantitative analysis of expected utility in the hands of users

Non-Clinical Performance Testing

- Technical performance of the AI system by itself
  - Interaction between the AI system and the user in the intended manner is not necessary
  - Can be component-by-component

Non-Clinical Testing

- Some potential non-clinical testing questions for an ultrasound image guidance device
  - Is the AI system able to assess/determine
    - Image quality
    - Closeness of the probe to the desired location to acquire a particular US view
    - Desired manipulation of the probe that will improve image quality
    - …

Basic Steps in Non-Clinical Testing

- Define
  1. Non-clinical testing questions to be addressed
  2. Assessment method and metric(s)
  3. Reference standard or comparator (ideally, “ground truth”)
  4. Data set
- Provide relevant statistical data on the assessment results, sub-group analyses

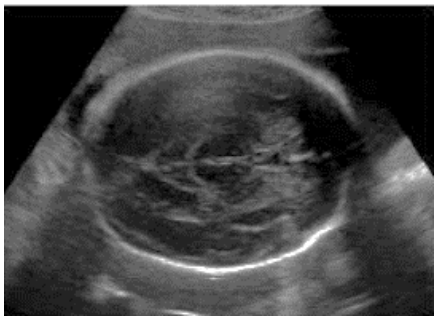
Example (Assessment Method & Metric)

- Non-clinical testing question
  - Algorithm performance in assessing image quality
- Assessment method:
  - Compare the quality ranking for pairs of images between
    - Algorithm
    - Experts
- Assessment metric:
  - Concordance measured by Kendall's τ

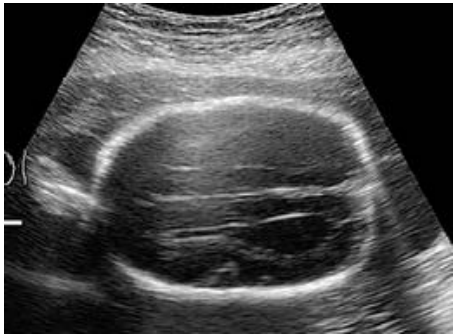
Example (Reference Standard or Comparator)

- Comparator
  - Side-by-side comparison of two US images by experts in image interpretation

I_1



I_2



Computer	$I_1 > I_2$	
Expert 1	$I_1 > I_2$	Concordant: C
Expert 2	$I_1 < I_2$	Discordant: D
Expert 3	$I_1 > I_2$	Concordant: C
Expert 4	$I_1 > I_2$	Concordant: C

$$\tau = \frac{N_C - N_D}{N_C + N_D}$$

Example (Data Set)

- Independent from the data set used for algorithm training
- Consideration for
  - Relevance and representativeness of the data set for the component to be tested
  - Sample size

Example (Results)

- 95% confidence interval for concordance between algorithm and experts
- Relevant comparisons
- Relevant sub-group analyses
- …

Clinical Performance Testing

- Clinical safety and effectiveness of the guidance device for its intended use, when used by the intended user
  - Clinical testing
  - Accepted virtual/physical systems designed to capture clinical variability
  - Comparison to a closely-related device with established clinical performance
  - Other sources that are appropriately justified

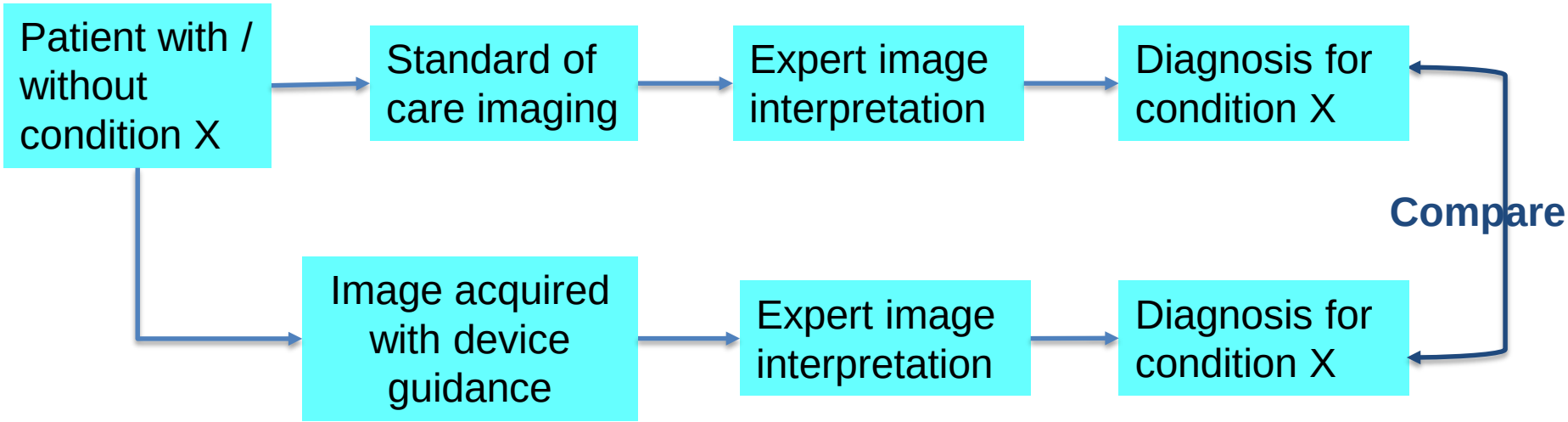
Clinical Performance Testing

- A quantitative evaluation of
  - Diagnostic utility and quality of images acquired / optimized
  - Performance in representative user and patient populations, under anticipated conditions of use

Example

- Comparison of diagnostic performance of experts, between
  - Images acquired with standard of care
  - Images acquired by a representative user population with guidance from the AI algorithm

Example (Continued)



Clinical Performance Testing

- Appropriateness of the clinical testing method depends on device benefits/risks
  - Image quality assessment comparison
    - Standard of care images
    - Acquired with the help of device
  - Comparison to a closely-related device with established clinical performance
  - Meeting a benchmark
  - …
- Topic of discussion at a Q-Sub meeting

Statistical Analysis Plan

- Pre-specification!
- Some elements (not exhaustive)
  - Endpoints
  - Statistical analysis methods
  - Process for defining truth (reference standard)
  - Statistical and clinical justification of sample size
  - Plan for multiple hypothesis testing if appropriate
  - Plan for handling missing data

Statistical Analysis for Sub-Groups

- Sufficient number of cases from important sub-groups to estimate performance and confidence intervals
- Powering for sub-groups may not be necessary
  - Depending on the clinical application
  - Unless specific sub-group performance claims are included

Technological Characteristics

- Needed for
  - Understanding scope of a change in a modification
  - Comparing two devices
  - May reduce performance testing requirements
- Flowchart describing processing, inputs, outputs, features, models, classifiers
- Algorithm training methods and training **DATA SETS**
- Algorithm parameter selection

Algorithm Modifications

- Potential to discuss plans for modifications during the initial premarket review
- Pre-determined change control plan
  - Ability to manage/control resultant risks of modifications

Generalizability

- Heightened importance because of the data-driven nature of most AI algorithms
- In addition to patient and user populations, how does the performance generalize to
  - Data acquisition equipment / conditions
  - Data acquisition sites
  - …


FDA
U.S. FOOD & DRUG
ADMINISTRATION