COLLEGE of AMERICAN PATHOLOGISTS

A proposed framework for deploying AI/ML in the clinical laboratory

Jansen N Seheult, *MB BCh BAO, MSc, MS, MD, FCAP*
Member, Machine Learning Working Group
Member, Artificial Intelligence Committee

Senior Associate Consultant & Assistant Professor,
Divisions of Hematopathology and Computational Pathology & AI,
Department of Laboratory Medicine and Pathology,
Mayo Clinic - Rochester

June 22, 2022

Terminology

- Verification: The process by which a laboratory determines that an unmodified FDA-cleared/approved test performs according to the specifications set forth by the manufacturer when used as directed
- Validation: The process used to confirm with objective evidence that a laboratory-developed or modified FDA-cleared/approved test method or instrument system delivers accurate and reliable results for the intended application

Understanding the Decision Summary

- Intended use – claimed model purpose
- Indications for use – A general description of the disease or condition the device will diagnose, treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended. Any differences related to gender, race/ethnicity, etc. should be included in the labeling.
- Sample size and distribution of data used in the model training and validation
- Inclusion and exclusion criteria
- How “ground truth’ was determined
- Performance claims
- Data compatibility, including how missing data are handled

Case Study: Prostate biopsy WSI analysis

- **FDA approved - For in vitro diagnostic (IVD) use only**
- **Indications for use:**
  - **Software only device intended to assist pathologists in the detection of foci that are suspicious for cancer during the review of scanned whole slide images (WSI) from prostate needle biopsies prepared from hematoxylin & eosin (H&E) stained formalin-fixed paraffin embedded (FFPE) tissue.**
  - **The software is intended to be used with slide images digitized with Scanner X and visualized with Vendor X's WSI viewing software.**
  - **The software is an adjunctive computer-assisted methodology and its output should not be used as the primary diagnosis. Pathologists should only use the software in conjunction with their complete standard of care evaluation of the slide image.**

Case Study: Prostate biopsy WSI analysis

- Training dataset: ~35,000 de-identified slides from single US laboratory between 2013-2017 and imaged with scanner Y
- Tuning dataset: ~6,000 slides prepared and stained at a single site and imaged with Scanner Y
- **Test datasets:**
  - **Tuning dataset (~6,000 slides) imaged with Scanner X**
  - **~11,000 slides prepared at >200 external sites but diagnosed at internal site and imaged with Scanner Y**
- **Approximately 80% of slides in training, tuning and testing datasets were collected from Caucasian patients, with approximately 8-9% from Black/ African American patients and 3% from Asians.**

Case Study: Accuracy characteristics

- **Accuracy study: Cancer (n = 311) and Benign (n = 417)**
  - Sensitivity = 94.5% (95% CI: 91.4 – 96.6%)
  - Specificity = 94.0% (95% CI: 91.3 – 95.9%)
  - Accuracy = 94.2% (95% CI: 92.3 – 95.7%)
- **Clinical study: Cancer (n = 171) and Benign (n = 356) read by 16 pathologists**
  - Assisted macro-averaged sensitivity = 96.8%
  - Assisted macro-averaged specificity = 89.5%
- **Case breakdown:**
  - Cancer: ~50% had tumor size ≤ 0.5mm & 50% > 0.5mm, ~2% with PIN, ~3-4% ASAP
  - Benign: ~88% without atrophy, PIN or treatment effects

Case Study: Precision characteristics

- Cancer (n = 35) and Benign (n = 36)
- Within-scanner: Slides scanned three times (3 reps) using one scanner/ operator
  - Cancer: 99.0% (95% CI: 94.8 – 99.8%) of all scans and 97.1% (34/35) of all slides produced correct results
  - Benign: 94.4% (95%CI: 88.4 – 97.4%) of all scans and 88.9% (32/36) of all slides produced correct results
- Reproducibility: Slides scanned once with three different scanners at different locations and by three different operators (one operator per scanner)
  - Cancer: 100.0% (95% CI: 96.5 – 100.0%) of all scans and 100.0% (35/35) of all slides produced correct results
  - Benign: 93.5% (95%CI: 87.2 – 96.8%) of all scans and 88.9% (30/36) of all slides produced correct results

Verifying manufacturer's accuracy claim

- H_0: Accuracy = P_0 versus H_1: Accuracy < P_0 (or Accuracy = P_1)
- With (1 – α)% confidence level and (1 – β)% power for detecting an effect of $P_1 - P_0$, the required sample size for cases is obtained from:

$$n = \frac{\left[Z_\alpha * \sqrt{P_0(1-P_0)} + Z_\beta * \sqrt{P_1(1-P_1)}\right]^2}{(P_1 - P_0)^2}$$

- For example, if the laboratory wishes to compare locally determined accuracy of a software or algorithm to the manufacturer's claim of 94.2%, the sample size required to have 95% confidence and 80% power to detect a difference of 5% from the claimed accuracy of 94.2% would be:

$$n = \frac{\left[1.645 * \sqrt{0.942(1-0.942)} + 0.84 * \sqrt{0.892(1-0.892)}\right]^2}{(0.892 - 0.942)^2} = 166$$

- To detect a difference of 10% from the claimed accuracy with 95% confidence and 80% power:

$$n = \frac{\left[1.645 * \sqrt{0.942(1-0.942)} + 0.84 * \sqrt{0.842(1-0.842)}\right]^2}{(0.842 - 0.942)^2} = 48$$

Verifying manufacturer's accuracy claim

- Dataset balance of cancer versus benign may influence choice of evaluation metric and required sample size
- How similar is your verification dataset to the manufacturer's accuracy and clinical study sets?
  - Your verification dataset should reflect your local patient population
  - Failure to verify manufacturer's stated claim may be driven by systematic differences in study sample characteristics

Verifying manufacturer's accuracy claim

- **Accuracy verification study performed with 166 samples (cancer = 83, benign = 83)**

$X^2 = 5.37$ $P = 0.15$	Vendor n = 728	Local lab n = 166
White	598	140
Black/ AA	58	18
Asian	22	3
Other	50	5

$X^2 = 7.69$ $P = 0.05$	Vendor n = 728	Local lab n = 166
Tumor ≤ 0.5mm	147	35
Tumor > 0.5mm	153	48
Benign (no atrophy/ PIN/ tx)	366	67
Other benign	51	16

- **Observed accuracy: 91.6% (152/166)**
- **Using one-sample test of proportion versus manufacturer's claim of 94.2%, p = 0.084 for one-sided alternate hypothesis**

Verifying manufacturer's precision claim

- **Simple precision**
  - **Repeatability: 10 slides (1:1 ratio of cancer: benign) scanned three times (3 reps) using one scanner/ operator**
  - **Reproducibility: 10 slides (1:1 ratio of cancer: benign) scanned once with different scanners, at different locations, by different operators (as appropriate)**
  - **Two-sample test of proportions:**
    - **Repeatability for cancer: 96.7% (95% CI: 83.3 – 99.4%) of 30 local scans compared with 99.0% (95% CI: 94.8 – 99.8%) of manufacturer's 105 scans (test if observed proportion significantly lower than manufacturer's claim: $p = 0.178$ at $\alpha = 0.05$)**
    - **Repeatability for benign: 86.7% (95%CI: 70.3 – 94.7%) of 30 local scans compared with 94.4% (95%CI: 88.4 – 97.4%) of manufacturer's 108 scans (test if observed proportion significantly lower than manufacturer's claim: $p = 0.075$ at $\alpha = 0.05$)**
- **Complex precision (ISO 16140)**

Case Study: What's next?

- Verification of accuracy and precision claims are not the end of your responsibilities as a Laboratory Director
  - Think PARR for method verification: Precision, Accuracy, Reportable range, Reference interval
- Try to break the model to understand its limitations
- Equipment qualification

COLLEGE of AMERICAN PATHOLOGISTS