# *Statistical Insight on pre-IDE*

OIVD Submissions Workshop

Marina V. Kondratovich, Ph.D.
Associate Director for Clinical Studies
OIVD, CDRH, FDA

April 27, 2011

# Outline

I. Introduction

II. Biases in clinical study:
selection bias;
spectrum bias;
verification bias;
cutoff selection bias

III. Patient Specific Score

## Key Elements

- Intended Use (IU)
  What is device supposed to do?

- Indications for Use (IFU)
  When should it be used?

- Both analytical and clinical data are supporting evidence for Intended Use and Indications For Use

# Intended Use Statement (how/by whom device is used)

- What is the device measuring, identifying or detecting? (analyte, organism, .. )
- Specimen types, matrix (whole blood, serum,..)
- Conditions for use (hospital lab, home use,..)
- What type of data output? (quantitative, qualitative, semi-quantitative)

# Indication for Use Statement (for what/on whom device is used)

- Target condition

- a particular disease, a disease stage, health status, or any other identifiable condition of event within a patient

- Target population (intended use population)

- those subjects for whom the test is intended to be used

- Medical Testing Contexts

- as, for screening, diagnosis, monitoring, prognosis, etc.

# *Examples of Medical Testing Contexts for cancer IVDs*

- **Diagnosis** (target condition is present or not during the time of testing);
- **Screening** (maybe in a general population (asymptomatic subjects at average risk) or a subpopulation (subjects at high risk)
- **Risk assessment** (assessment of predisposition to disease in future)
- **Prognosis** (stratifying already diagnosed cancer patients into poor or good prognosis)
- **Monitoring** (is therapy working for a patient?)
- **Companion Diagnostics/Co-development paradigm** (Therapeutic response prediction)

* This is not a comprehensive list

# Intended Use/Indication For Use

Example 1:

The HPV HR test is an *in vitro* diagnostic test for the qualitative detection of DNA from 14 high-risk Human Papilloma Virus (HPV) types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, and 68) in cervical specimens. *To screen* patients with atypical squamous cells of undetermined significance (ASCUS) cervical cytology results *to determine the need for referral to colposcopy.*

# Example 2

The OVA1 Test is a qualitative serum test that combines the results of five immunoassays into a single numerical score. It is indicated for women who meet the following criteria: over age 18; ovarian adnexal mass present for which surgery is planned, and not yet referred to an oncologist. *The OVA1 Test is an aid to further assess the likelihood that malignancy is present when the physician's independent clinical and radiological evaluation does not indicate malignancy.* The test is not intended as a screening or stand-alone diagnostic assay.

# CLSI documents helpful for analytical studies

| | |
|---|---|
| EP05-A2 | Precision |
| EP06-A | Linearity |
| EP07-A2 | Interference |
| EP09-A2 | Comparison studies |
| EP12-A2 | Qualitative tests |
| EP17-A | Limit of detection and limit of quantitation |
| EP21-A | Total error |
| EP25-A | Reagent stability |
| C28-A3 | Reference intervals |
| MM17-A | Multiplex |

## Intended Use/Indication For Use drives:

- Study design
- Study should match intended use
- Kinds of patients (Asymptomatic,..)
- Clinical sites (e.g. doctor's office, ER, hospital)
- Sample size justification
- Clinical usefulness

Devices are regulated by their intended use:

Total PSA for diagnosis – PMA

Total PSA for monitoring already diagnosed patients – 510(k)

# N subjects in the clinical study
(N subjects from target population)

Every subject

## Candidate Test:

Positive,
Negative

## Clinical Reference Standard (Gold Standard):

D+ = Target condition present,
D- =Target condition absent

| | | Colpo/Biopsy | | Total |
|---|---|---|---|---|
| | | CIN2+ | Not-CIN2+ | |
| HPV HR | Pos | 64 | 693 | 757 |
| | Neg | 5 | 550 | 555 |
| Total | | 69 | 1243 | 1312 |

Clinical performance refers to the degree of agreement between the results of the Candidate test and the results from the Clinical Reference Standard (CRS), “Gold” Standard.

# Candidate Test

- Finalize assay steps before the pivotal clinical study
- Define interpretations of all outputs, including equivocal

Example:

S/Co ≤ 1.0, Negative;
S/Co > 1.0, Positive

Example:

S/Co ≤ 0.9, Negative;
0.9< S/Co ≤1.1, Equivocal;
S/Co > 1.1, Positive

Invalid result (control failed) ≠ Equivocal

All results should be reported

# Clinical Reference Standard

*Clinical Reference Standard, CRS* (Gold Standard)-
best available method for establishing the presence or absence of the target condition
(for example, colposcopy/biopsy for cervical cancer)

- ❑ Target condition is not necessary a disease (for example, it can be a success of some treatment)
- ❑ Target condition can be present at the same time when test T is applied; it can be present in future.

Basic principles:

1) Candidate test results CANNOT be used in CRS
2) CRS can classify <u>each</u> subject from the target population as “Target condition present” or “Target condition absent”.

# Clinical Protocol for Pivotal Study

- Consistent with intended use
- Site types (e.g., Point of Care)
- Inclusion/exclusion criteria
- Clinical reference standard
- Clinical performance measures
- Performance goals
- Statistical methodology
- Sample size

# Banked (retrospective) samples

## A good reason for pre-IDE

May be allowed

- How representative are banked samples (inclusion/exclusion criteria)
- Clinical context on specimens
- Only leftovers from big tumors (sample volumes)?
- Storage does not impact analyte of interest

**Provide unbiased estimates of performance**

# Clinical Performance Characteristics

- Clinical sensitivity, clinical specificity
- Positive and negative likelihood ratios
- Positive and negative predictive values along with prevalence

| | | Colpo/Biopsy | | Total |
|---|---|---|---|---|
| | | CIN2+ | Not-CIN2+ | |
| HPV HR | Pos | 64 | 693 | 757 |
| | Neg | 5 | 550 | 555 |
| Total | | 69 | 1243 | 1312 |

| | | Colpo/Biopsy | | Total |
|---|---|---|---|---|
| | | CIN2+ | Not-CIN2+ | |
| HPV HR | Pos | 64 | 693 | 757 |
| | Neg | 5 | 550 | 555 |
| Total | | 69 | 1243 | 1312 |

Sensitivity = **92.8%** (64/69)
Specificity = **44.3%** (550/1243)

Risk of D+ for T+ (PPV) = **8.5%** (64/757)
Risk of D+ for T- (=1-NPV) = **0.9%** (5/555)
Negative predictive value (NPV) =99.1% (550/555)
Prevalence (pre-test risk of disease) = **5.3%** (69/1312)

Positive likelihood ratio=se/(1-sp) = **1.66** (92.8%/55.7%)
Negative likelihood ratio=(1-se)/sp=**0.16** (7.2%/44.3%)=1/6.1

# *Estimation*

❑Point estimates and 95% confidence intervals

Sensitivity = 92.8% (64/69) with 95% CI: 84.1% to 96.9%

❑ Statistical: 95% lower limit is above 84% - we can say sensitivity is significantly above 84% (p-value <0.05)

❑ Diagnostic devices: **prefer estimation** to hypothesis testing (p-values)

Estimation: point estimate and 95% confidence interval

Confidence interval is wide with small sample size

"Right" subjects in the clinical study

Simply increasing the overall number of subjects in the study will do nothing to reduce bias.

We considered an ideal scenario when N randomly selected subjects are from the intended use population and each subject has result of the test and verification of disease (D+, D-).

Potential Biases

1) **Selection bias** (when the study population does not represent the IU population)
2) **Spectrum bias**
3) **Verification bias**
4) **Cutoff selection bias**

# Example 1 of <u>inappropriate</u> study design (selection bias)

Cervical cancer: $T_{Old}$ – HPV test used in current practice to make a decision about need for colposcopy

$T_{New}$ – new HPV test.

Only subjects referred to colposcopy were included in the clinical study.

| $T_{New}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **A** | **B** | |
| Neg | **C** | **D** | |
| | | | N |

**Disease D+**

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | $a_1$ | | |
| Neg | $c_1$ | | |
| | $a_1+c_1$ | | |

**Non-Disease D-**

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | $a_0$ | | |
| Neg | $c_0$ | | |
| | $a_0+c_0$ | | |

# 900 subjects from intended use population

**Disease D+**

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **60** | 7 | 67 |
| Neg | **7** | 1 | 8 |
| | **67** | 8 | 75 |

**Non-Disease D-**

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **290** | 10 | 300 |
| Neg | **10** | 515 | 525 |
| | **300** | 525 | 825 |

Unbiased: $Se_{New}$ = 89.3% (67/75) $Sp_{New}$ = 63.6% (525/825)

- Only subjects positive by $HPV_{Old}$ were included in study (367 subjects)
- Biased estimation of $T_{New}$ performance:

$Se_{New}$ = 89.5% (60/67); $Sp_{New}$=3.3% (10/300) →

It appears that $T_{New}$ is not informative (random test).

# Example 2 of inappropriate study design (selection bias)

| Current Practice<br>Pre-Surgical Assessment by Physician | | |
|---|---|---|
| **Malignant** | **Non-Malignant** | |
| All subjects were referred to oncology centers | Subjects were operated in oncology centers | Subjects were operated in places other than oncology centers |
| **Subjects of the Clinical Study** | | |

If one include only subjects from oncology centers in the clinical study, then this study will have selection bias.

# Example 3 of inappropriate study design (selection bias)

- Alzheimer's disease

In the study, the subjects with severe AD and healthy subjects were included => Selection bias – overestimation of performance.

- If the healthy subjects are not part of intended use population, do not include them in the clinical study (overestimation of specificity).

- Healthy subjects are used for determination of reference intervals.

# 2) Spectrum Bias

***Example***

Diseased subjects in the Intended Use population =
50% of Stage II and 50% of Stage I
Test ABC has sensitivity for Stage II = 90%;
Stage I = 50%

Sensitivity of test ABC in the IU population =
0.5 * 90%+ 0.5 * 50% = **70%**

*Retrospective samples* in the clinical study
80% of Stage II and 20% of Stage I:
Sensitivity in the clinical study =0.8 * 90% + 0.2 * 50* = **84%**

***Sensitivity is biased (overestimated)***

# 3). Verification Bias

❑ ***We know that some clinical reference standards are expensive or invasive: it may be impossible, or even unethical, to apply the clinical reference standard to <u>all</u> clinical study subjects.***

Examples

• Claims related to screening;

• Test under investigation is applied to in vitro samples and the clinical reference standard is applied to human subjects.

# 3). Verification Bias

***Example***

Clinical study with 100 subjects: each subject has verification of disease and test result

| | | Gold Standard | | Total |
|---|---|---|---|---|
| | | D+ | D- | |
| Test | Pos | 20 | 5 | 25 |
| | Neg | 30 | 45 | 75 |
| Total | | 50 | 50 | 100 |

Se = 40% (20/50)

Sp = 90% (45/50)

***Example*** (cont.)

Subjects were referred to the CRS based on the “Current clinical practice”.

In the study, all 25 subjects with pos. test results -> CRS; only 1/3 of 75 subjects with neg. test results -> CRS.

Analysis of the data with verified disease status

| | | CRS | | Total |
|---|---|---|---|---|
| | | D+ | D- | |
| Test | Pos | 20 | 5 | 25 |
| | Neg | 10 | 15 | 25 |
| Total | | 30 | 20 | 50 |

Se = 67% (20/30)

Sp = 75% (15/20)

Sensitivity is biased (overestimated)

Specificity is biased (underestimated)

*Verification Bias* occurs when a non-random group of subjects in the clinical study selectively receive clinical reference standard.

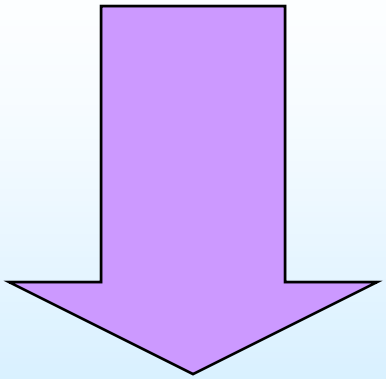
Prostate cancer

$T_{New}$ – new biomarker as an aid to make a decision who needs a prostate biopsy

Complex pattern describes how subjects are referred to prostate biopsy (current practice uses age, race, digital rectal exam, free PSA, family history etc)

How to evaluate $T_{New}$ in unbiased way?
Very challenging problem!

Some appropriate study designs where not ALL subjects have verified disease status
( three examples)



# Example I

Pap test

$T_{Old}$ – reading of a slide in laboratory by the current method (manual)

$T_{New}$ – computer-aided reading of the slide in laboratory

All slides positive either by $T_{Old}$ or by $T_{New}$ are referred to the Clinical Reference Standard (reading of a slide by adjudication committee, cytology truth);

A random sample of subjects with both negative test results (5-10%) are referred to the Clinical Reference Standard.

| | | $T_{Old}$ Pos | $T_{Old}$ Neg | |
|---|---|---|---|---|
| $T_{New}$ | Pos | **400** | **200** | |
| | Neg | **100** | **9,300** | |
| | | | | 10,000 |

| $T_{New}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **400** | **200** | |
| Neg | **100** | **9,300** | |
| | | | N |

**Clinical Reference Standard+**

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **380** | **100** | |
| Neg | **60** | **[30]** | |
| | | | **[$N_1$]** |

**Clinical Reference Standard -**

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **20** | **100** | |
| Neg | **40** | **[900]** | |
| | | | **[$N_0$]** |

10% of 9,300 slides=930. Among them, 30 have CRS+ and 900 have CRS-.

**The unbiased estimates of sensitivities and specificities for $T_{Old}$ and $T_{New}$ can be constructed (multiple imputation).**

# *Example II*

- All subjects which are positive either by $T_{Old}$ or by $T_{New}$ are referred to Clinical Reference Standard (CRS);
- No subjects with negative on both tests are referred to CRS.

| $T_{New}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **A** | **B** | |
| Neg | **C** | **D** | |
| | | | N |

**Disease D+** **Non-Disease D-**

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | $\mathbf{a_1}$ | $\mathbf{b_1}$ | |
| Neg | $\mathbf{c_1}$ | | |
| | | | |

| $T_{new}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | $\mathbf{a_0}$ | $\mathbf{b_0}$ | |
| Neg | $\mathbf{c_0}$ | | |
| | | | |

| $T_{New}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **200** | **200** | 400 |
| Neg | **50** | **9550** | |
| | 250 | | 10,000 |

**Disease D+**

| $T_{New}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **140** | **140** | **280** |
| Neg | **20** | | |
| | **160** | | |

**Non-Disease D-**

| $T_{New}$ \ $T_{Old}$ | Pos | Neg | |
|---|---|---|---|
| Pos | **60** | **60** | **120** |
| Neg | **30** | | |
| | **90** | | |

Ratio of sensitivities (TPR) can be estimated

$$\hat{TPR}_2 / \hat{TPR}_1 = 280/160 = 1.75$$

Ratio of 1-specificity (FPR) can be estimated

$$\hat{FPR}_2 / \hat{FPR}_1 = 120/90 = 1.33$$

It is possible to evaluate performance of $T_{New}$ if

- There is an increase in TP rates
  $TPR_{New}/TPR_{Old} > 1$
  (1.75 in the example)

- The increase in TP rates is larger than the increase in FP rates
  $TPR_{New}/TPR_{Old} > FPR_{New}/FPR_{Old}$
  (1.75 > 1.33 in the example)

$T_{New}$ has higher PPV and higher NPV than $T_{Old}$.

* For details, see Kondratovich, M.V (2008) Comparing Two Medical Tests When Results of Reference Standard Are Unavailable for Those Negative via Both Tests, *Journal of Biopharmaceutical Statistics,* 18: 1; 145-166

## *Example III* Studies with Follow-Up

- New prognostic biomarker (positive/negative)

Subjects with positive biomarker have a higher risk of disease, for example, in next 5 years than those with negative biomarker.
Example: cervical cancer; breast cancer.

- Clinical study with follow-up; all subjects have annual visits.

- Possible scenario: at every annual visit, only subjects who are positive by current practice have a formal verification of disease status.

For example, cervical cancer: only subjects with Pap abnormal and/or HPV positive results or other risk factors are referred to colposcopy (clinical reference standard);
The data of the clinical study can have potential verification bias.

- If the Biomarker is correlated with the “Procedure of Referral to Clinical Reference Standard”, then estimations of
  - Risks are biased;
  - Relative risk (RR) is biased.

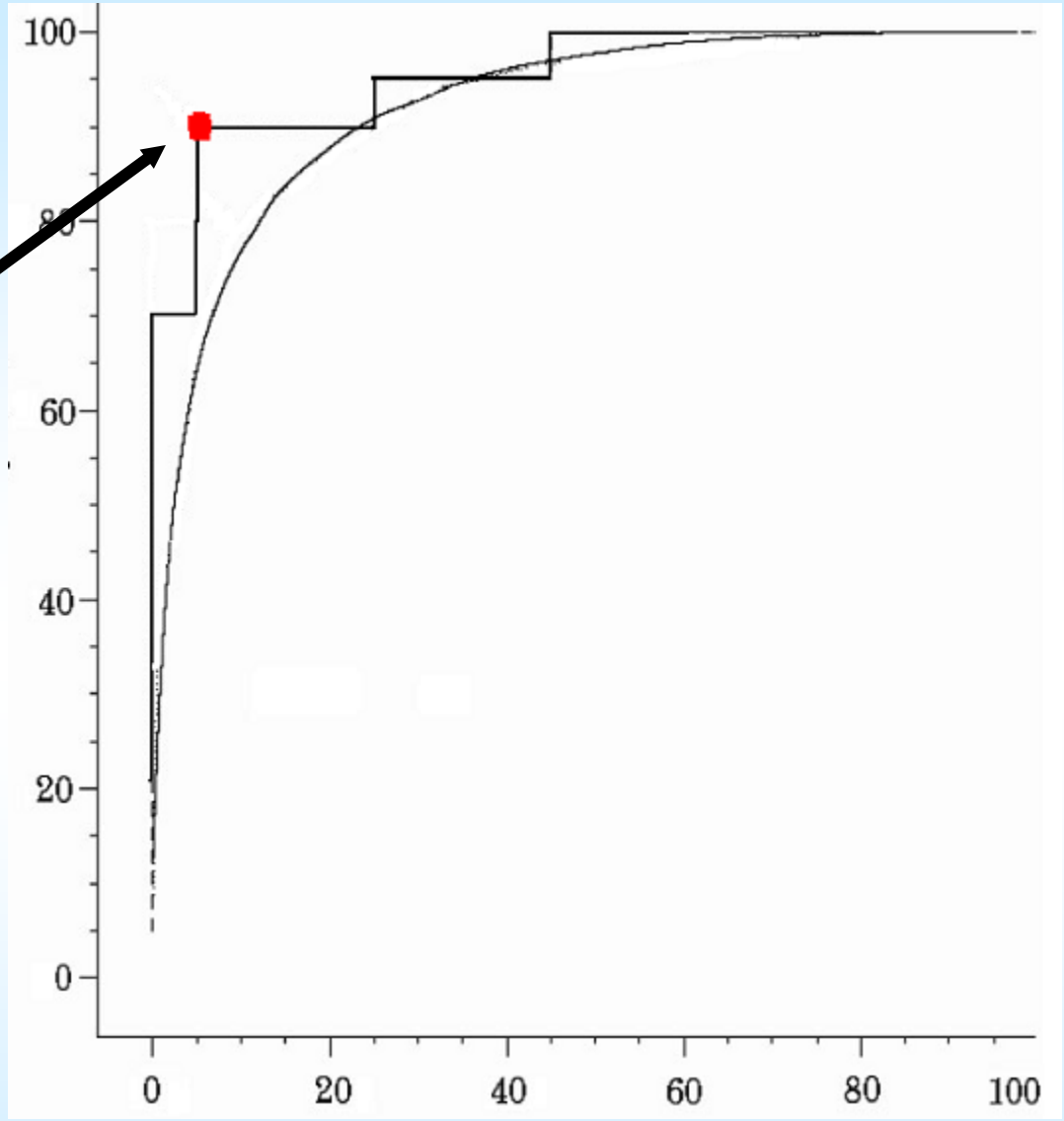
- If the Biomarker is not correlated with the “Procedure of Referral to Clinical Reference Standard”, then estimations of
  - Risks are biased;
    Underestimated (less than sensitivity of the procedure of referral to CRS);
    The more sensitive the procedure of referral to CRS, the better (less bias in risk estimation)
  - Relative risk (RR) is unbiased.

# 4). Selection of Cutoff

❑***Classical Approach:***
Cutoff is selected BEFORE the pivotal study
(based on analytical studies, pilot data, convenience samples and so on).
Then this cutoff is applied in the pivotal clinical study.

❑ Sometimes there may be little information available in the early phases of test evaluation => sometimes cutoff is selected in the pivotal study.
If cutoff is selected in the pivotal study as max of (sensitivity + specificity)); then this leads to too optimistic measures of clinical performance => an independent study is needed.

Overestimated Sensitivity and Specificity*




*Linnet K., Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin. Chem*. 1986; 32: 1341-6

- ***Different Approach***:

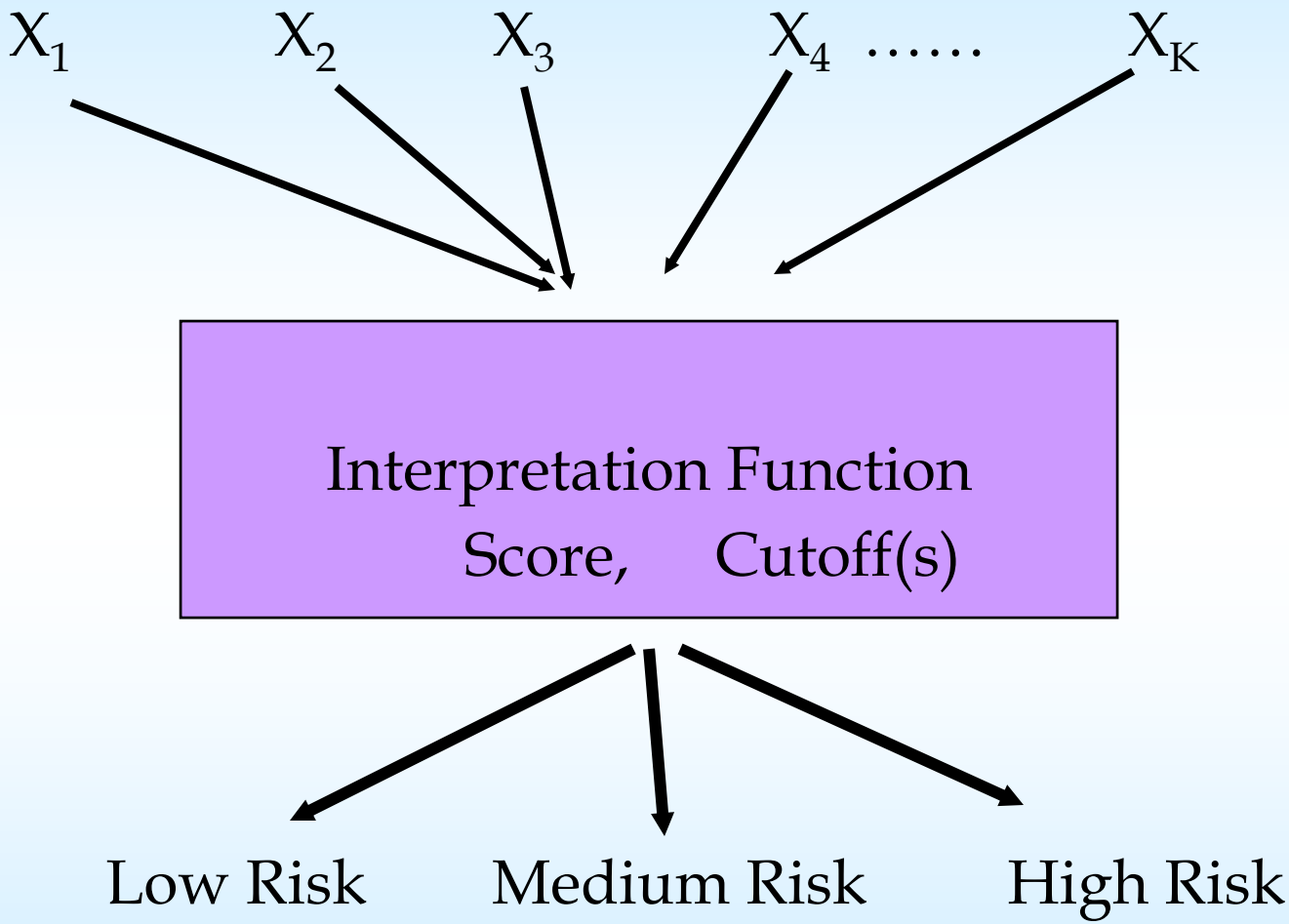
  Consider that level of sensitivity (or specificity) is pre-specified. In the pivotal study, cutoff is selected as an unbiased estimate of a corresponding percentile. Then no bias in estimation of clinical performance*.

  Confidence intervals around sensitivity and specificity will increase.

1) Pre-specify level of sensitivity (or specificity)
2) Use the same pivotal study for selection of the cutoff (as corresponding percentile) and estimation of clinical performance
3) CI is wider (use bootstrap)

* Kondratovich M, Yousef WA. Evaluation of accuracy and 'optimal' cutoff of diagnostic devices in the same study. Joint Statistical Meeting. 2005. ASA Section on Statistics in Epidemiology.

# *III. Patient Specific Score*




Combines the values of multiple variables using an interpretation function to yield a single, patient-specific result (e.g, a "classification", "score", "index", etc)

# *Training and Validation Steps*

- Training sets, testing sets
- Develop classifier
- Internal validation (cross validation)
- Lock classifier (interpretation function)
- Two different approaches for cutoff
  - select a cutoff in the training set or
  - pre-specified level of sensitivity (or specificity) ->cutoff will be selected in the pivotal study

# Independent Validation

- Performance of the Score is evaluated in the independent validation study (pivotal study)

- Performance in label: from validation study (pivotal study)

- Pivotal study represents intended use population

# Summary

- It is important to understand potential sources of bias so they can be avoided or minimized.

  Note: Simply increasing the overall number of subjects in the study will do nothing to reduce bias.

- Selection bias and verification bias
- We discussed also different scenarios of studies which produced unbiased estimation of clinical performance.
- Cutoff of the assay

# *Thank you!*

*Marina.Kondratovich@fda.hhs.gov*