Analytical and Clinical Studies Design Considerations for IVDs

2019 OIR Submissions Workshop

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

April 10, 2019

Outline

I. Different types of in vitro clinical tests

II. Analytical performance: precision

III. Clinical Performance:
archived samples,
potential biases,
tests with multiple outputs

I. Different types of in vitro clinical tests

What Type are Device Outputs

How results of the device are reported to a physician?

Qualitative test: binary outputs or
with multiple outputs of nominal type

Qualitative test: with 2 outputs (negative, positive)

with multiple outcomes
(e.g. genotyping of HCV with multiple outputs as 1a, 1b, 2, 3, 4, 5 and 6)

Nominal

- Nominal refers to data such as names/categories (nominal=name).
- Example: five different genotypes. May have numbers assigned (not for arithmetic purpose).

What Type are Device Outputs

How results of the device are reported to a physician?

Quantitative test: The amount or concentration of an analyte is measured and expressed as a numerical quantity value in measurements units.

Quantitative

- Values that can be subtracted and can be divided:
  Total PSA: values 50, 100, 150 (units)

Linearity of the device should be evaluated.

What Type are Device Outputs

How results of the device are reported to a physician?

Semi-Quantitative test: test with ordinal outputs

Ordinal

- Ordinal refers to quantities that have an ordering – order matters but not the difference between values.

  Example: urine dipstick with outputs:

  neg, trace, 1+, 2+, 3+;.

*Semi-quantitative test: not well defined term

Multi-Analyte Assays with Algorithmic Analyses (MAAA)

Assays based on an individual analyte

Machine Learning–based devices:

multiple analytes,

other variables related to the patient clinical information,

imaging

A device that:

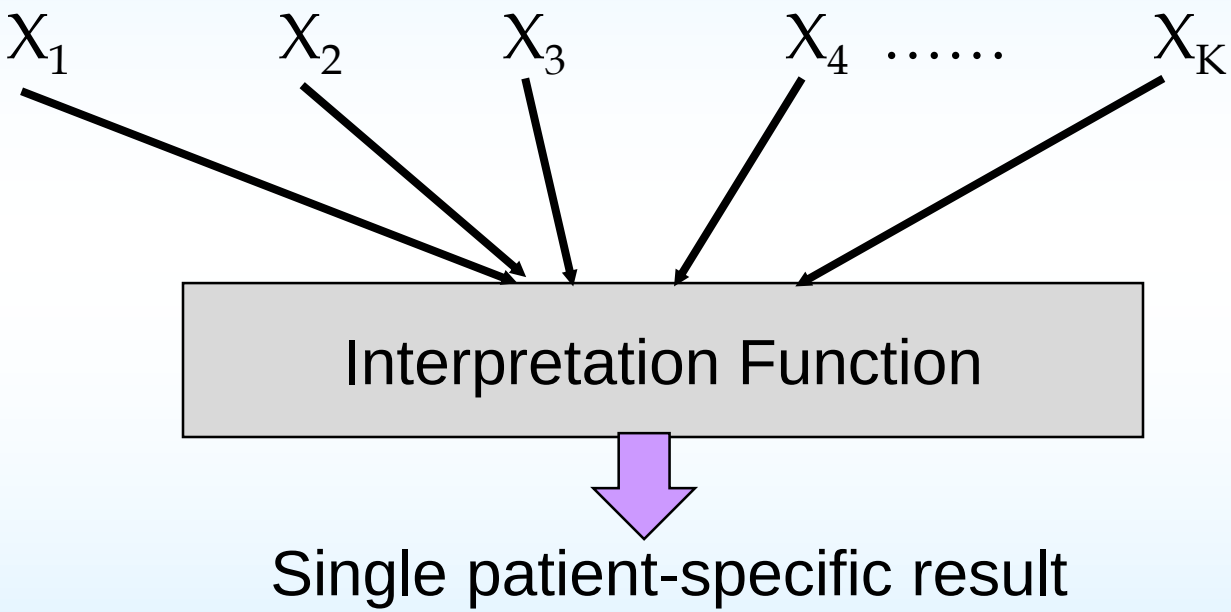
Combines the values of multiple variables using an interpretation function to yield a single, patient-specific result (e.g., “classification”, “score”, “index”).

This single patient-specific result can be

- “classification” (e.g., three classes: Normal, Prediabetic, Diabetic)
- numeric value (score, index

Multi-Analyte Assays with Algorithmic Analyses (MAAA)

- Analytes can be individually measured or in multiplex



Interpretation function: Logistic regression
Neural networks
Different methods

Multi-Analyte Assays with Algorithmic Analyses (MAAA)

Individual (6 variables)					
2 Covariates		**4 Biomarkers from whole blood**			
Race X_1	Family History X_2	X_3	X_4	X_5	X_6

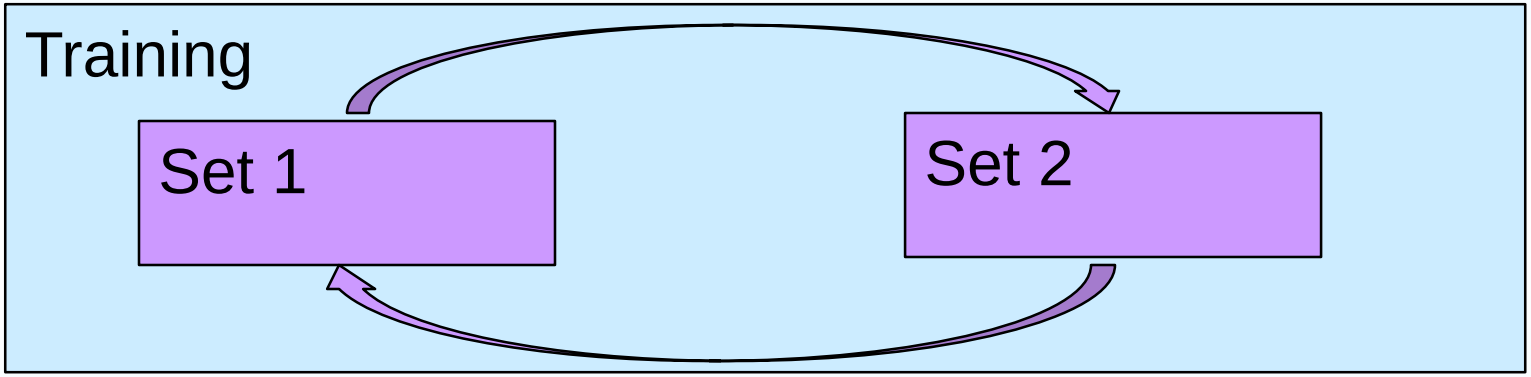
Score is a numeric value

- One cutoff for the numeric value of the score => Positive, Negative results (some similarities with binary qualitative test);
- Multiple cutoffs for the numeric value of the score => e.g., “low risk”, “medium risk”, “high risk” (some similarities with semi-quantitative test);

- Numeric values are probabilities (risks) of disease (risk assessment tests). It can be absolute risks, relative risks and other measures of the risk.

Multi-Analyte Assays with Algorithmic Analyses (MAAA)

Supervised Machine-Learning
A model used to distinguish “Diseased” vs “Non-Diseased” would be shown a data set with patients: 6 variables and **status of Disease** for each patient. This is called **“training”.**



Pivotal clinical study

In Vitro Clinical Test

Analytical performance (measuring device)	Clinical performance (related to the claim)
Analytical performance—does the test measure (detect) the analyte I think it does? Correctly? How reproducibly?	**Clinical performance**—is a patient test result associated with the expected clinical presentation of this patient?

Precision
Limit of Blank,
Limit of Detection,
Limit of Quantitation,
linearity,
……..

II. Analytical Performance: Precision

Precision Studies

Precision

closeness of agreement between … measured quantity values obtained by replicate measurements on the same .. objects under specified conditions.
NOTE: The 'specified conditions' can be, for example, repeatability conditions of measurement, intermediate precision conditions of measurement, or reproducibility conditions of measurement.

Repeatability

same lab,
same lot,
same operator,
same day,
same run

Within-Laboratory Precision

same lab,
different operators,
different days,
different runs,
different lots

Reproducibility

different labs,
different operators,
different days,
different runs

Intermediate Precision

Precision Studies

Example of reproducibility study

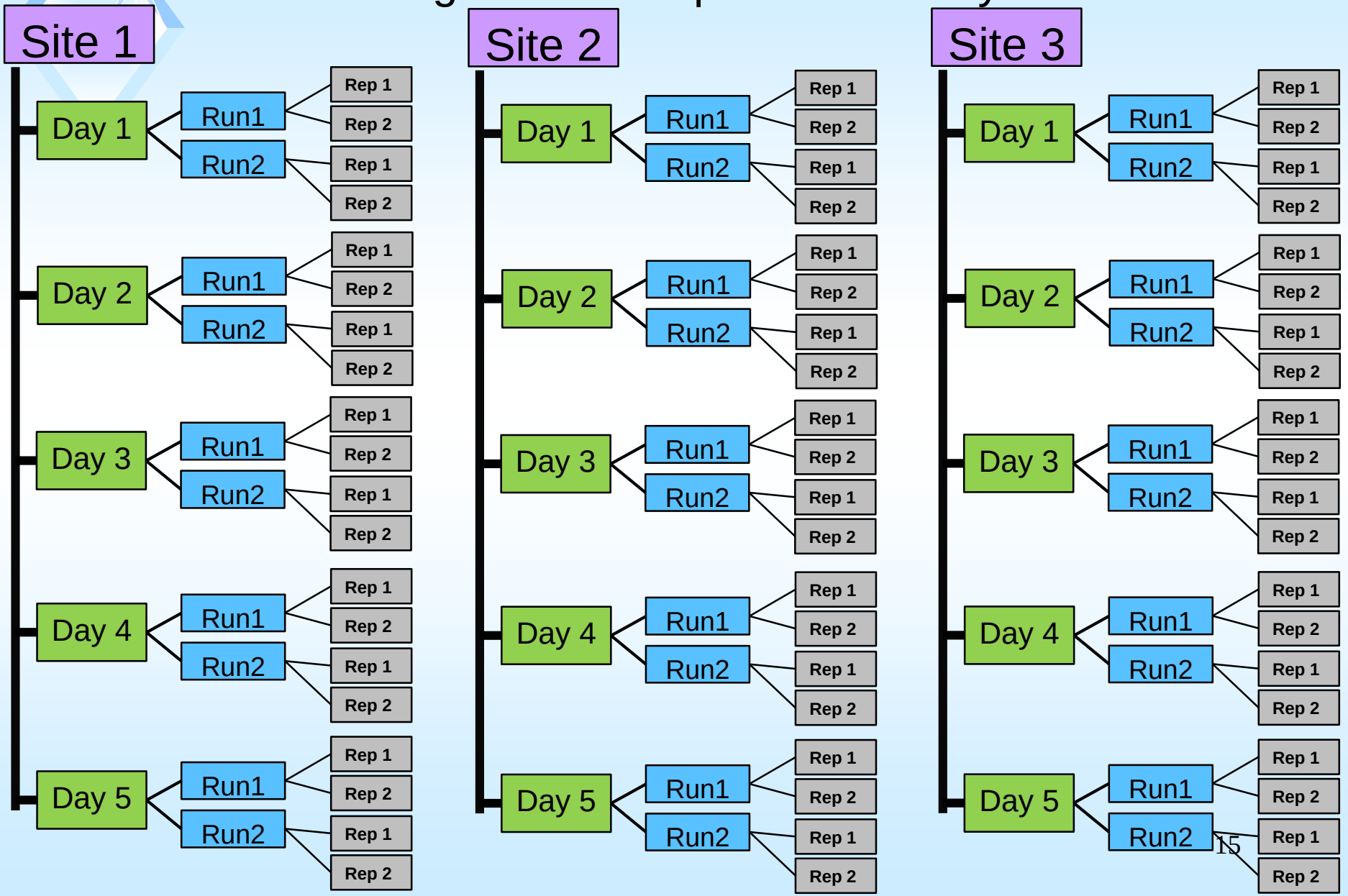
- 3 sites (1 internal + 2 external)
- 5 days per site
- 2 runs per day
- 2 replicate per run

❑Source of variability “operator-to-operator”

2 operators? Balanced design

Precision Studies

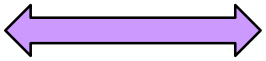
- Provide a diagram for the precision study



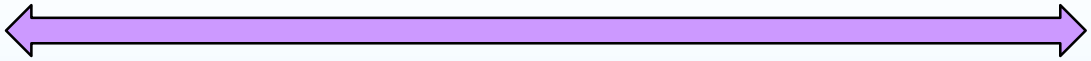
Precision Studies

For analysis of the data, use CLSI EP05-A3.

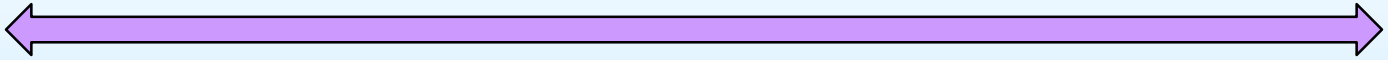
Reproducibility												
Mean	Repeatability (within-run)		Between-run		Between-day		Between-operator		Between-site		**Total**	
	SD	%CV	SD	%CV	SD	%CV	SD	%CV	SD	%CV	SD	**%CV**
….												



Repeatability



Within-Lab Precision



Reproducibility

Precision Studies

- **Source of variability "lot-to-lot"**

  Different study designs:

A) 3 sites

- each site has 3 lots
- 5 days
- each day

  2 runs with Lot 1,
  2 runs with Lot 2,
  2 runs with Lot 3
- each run has 2 replicates

Precision Studies

- **Source of variability "lot-to-lot"**

  Different study designs:

B) Two precision studies

Study 1

Evaluation of lot-to-lot precision at one site (usually internal)

Study 2

Reproducibility

3 sites but each site has the same lot

Through pre-Sub, discuss how different sources of imprecision will be evaluated (especially for specimens as fingerstick WB, saliva, fresh urine).

Precision Studies

Reproducibility incorporates lab-to-lab variability

- Usually it requires 3 different testing sites
- In some cases, it can be 3 different instruments within one site
  - device is minimally susceptible to environmental conditions
  - Variability because of operators skills is negligible

CLSI EP05-A3 (2014), section 4.2

Notes:

1) POC (including CLIA waivers (dual)) – 3 POC sites
2) Device will be run at a single site => no need in reproducibility study

Precision Studies for the Score

Consider two individual analytes X_1 and X_2,

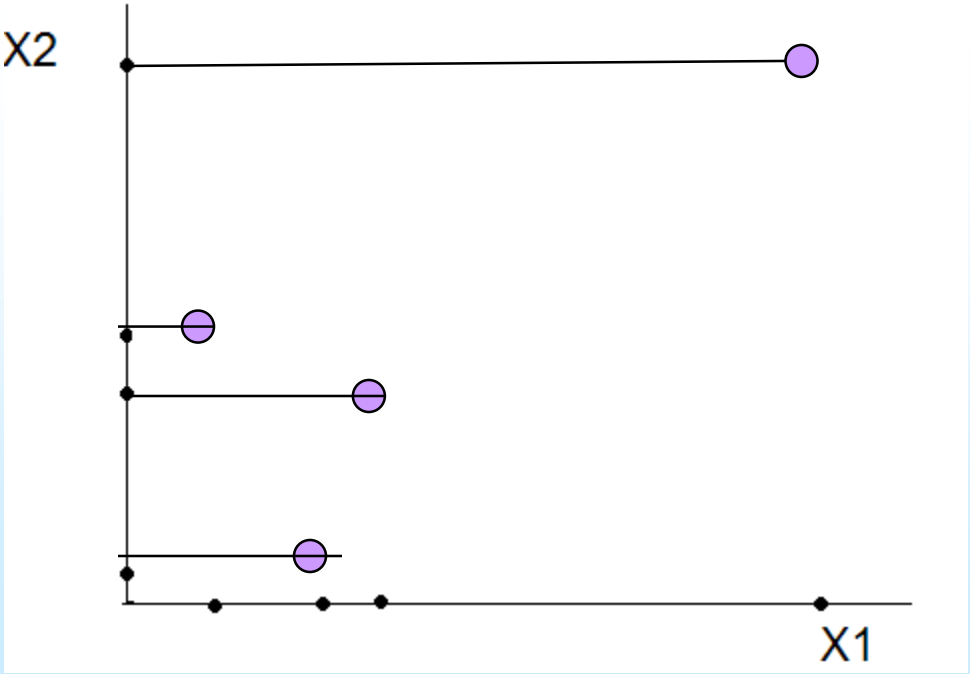
score is f(X_1, X_2).

Precision study includes **4 samples**.

Each sample has some particular value of X_1 and X_2

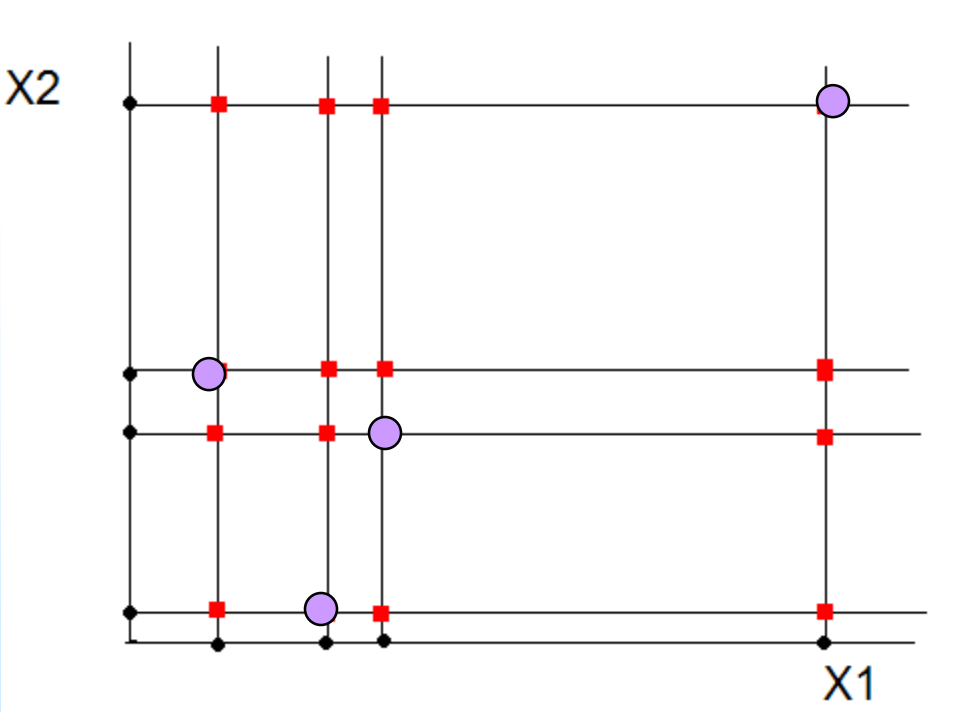
For each of 4 samples:

%CV of the Score, %CV for X_1, %CV for X_2



Precision Studies for the Score

- Score can be calculated at **16 points**
- Precision can be evaluated using **in silico** calculation (computer modeling) with normal distributions of measurement errors for each X_i
- Random measurement errors of analytes X_1, X_2, …, X_K are not correlated (because analytes are measured individually)

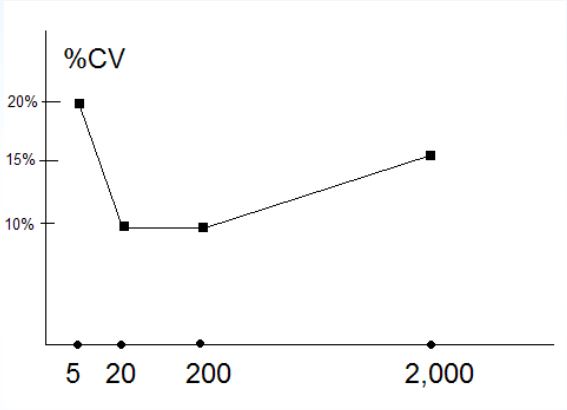


Precision Studies for the Score

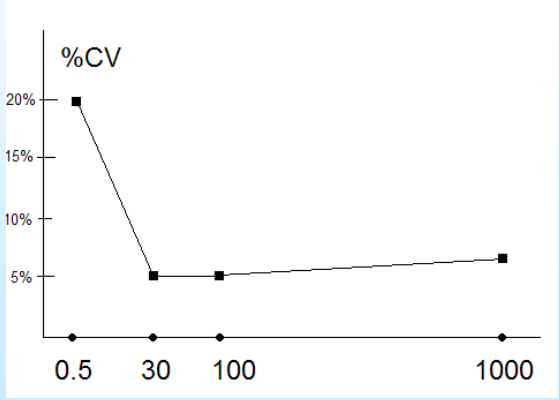
- Using precision profiles for X_1 and X_2, precision of the Score is calculated for the entire region of (X_1, X_2) values

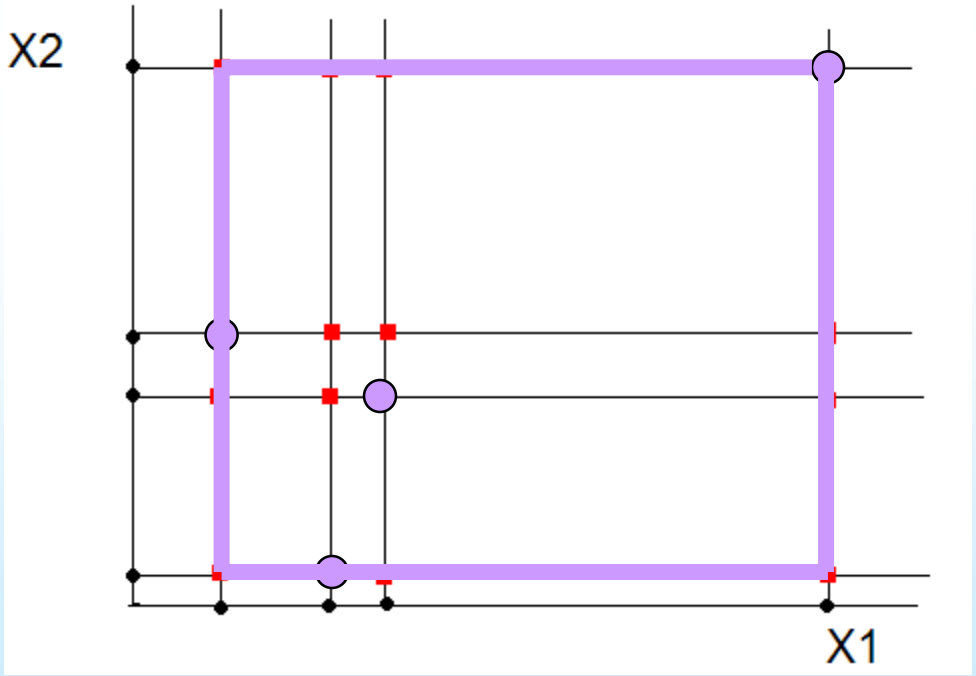
There are many possible combinations of the individual analytes amounts which give the **same value of the test score** and therefore, the samples with the same score but different combinations of the individual analyte amounts can **have different precisions.**

Precision profile for X_1



Precision profile for X_2





III. Clinical Performance

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.

Clinical Studies

Guidance for Industry, Clinical Investigators, Institutional Review Boards and Food and Drug Administration Staff –

Design Considerations for Pivotal Clinical Investigations for Medical Devices (2013)

The web address

http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm373750.htm

Section 8, pages 38-46

Qualitative Test

Two outputs: Pos, Neg

Analytical Accuracy

"**Analyte absent**" =
"Concentration of the analyte=0"
vs
"**Analyte present**"=
"Concentration of the analyte>0"

Reference method: The best available method for establishing the presence or absence of analyte

Recognized term:
ISO, VIM, CLSI, FDA guidance

Clinical Accuracy

"**Target condition absent**"

vs

"**Target condition present"**

The best available method for establishing the presence or absence of TC

Gold Standard,
Clinical reference standard,
Diagnostic accuracy criteria

No term in ISO, VIM

Do not use term "reference standard" for clinical accuracy.
It has different meaning, it is related to analytical performance:
VIM: measurement standard used for the calibration of working measurement standards in a given organization or at a given location

When “analytical accuracy” and “clinical accuracy” are the same concepts and when these are different concepts?

A) If Target Condition in clinical performance coincides with Presence or Absence of the analyte

“Target condition absent”=“Analyte absent”

Then Analytical accuracy=Clinical accuracy and
Reference method=Gold Standard

Examples: HIV, HCV, HBV, ..

B) If **“Target condition absent” ≠ “Analyte absent”**
then Analytical accuracy and Clinical accuracy are different concepts

Examples: 1) Analytical accuracy: “Mutation present” vs “Mutation absent”
Clinical accuracy: Target Condition=“Colon cancer in next 5 years”
2) Analytical accuracy:
“HPV present above threshold” vs “HPV absent or present below threshold”
Clinical accuracy: Target condition =“Cervical disease present”

Gold Standard for Target Condition

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

- Can be a single method or a combination of methods and techniques, including clinical follow-up.

- 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.

Archived samples

A good topic for pre-Sub

May be allowed for clinical study

- How representative are archived samples (inclusion/exclusion criteria)
  - Clinical context on specimens
  - Only leftovers from big tumors (sample volumes)? Re-testing of samples close to the cutoff (sample volume)?
- Storage does not impact analyte of interest

Basic principle:
Archived sample should provide unbiased estimates of test clinical performance.

Potential Biases

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**
2) **Spectrum bias**
3) **Verification bias**

1) Selection Bias

When the study population does not represent the IU population.

Examples of inappropriate study design

- Alzheimer's disease: intended use population=subjects with signs of memory loss.

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

Test ABC

Intended Use population		
Stage I	**50%**	**Sen=50%**
Stage II	**50%**	**Sen=90%**
Overall	**100%**	**70%** **0.5*50 + 0.5*90**

Archived Specimens		
Stage I	**20%**	**Sen=50%**
Stage II	**80%**	**Sen=90%**
Overall	**100%**	**82%** **0.2*50 + 0.8*90**

Sensitivity is biased (overestimated)

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)

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

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)

III. Clinical Performance: how to describe clinical performance

Consider Test with Two Outputs (Pos, Neg)

Let us have 1,300 subjects who are representative subjects from intended use population (target population). Each subject has results of the Test (Pos, Neg) and ("Gold Standard") (D+, D-).

		Colposcopy		
		D+	**D-**	Total
T	**Pos**	66	694	760
	Neg	4	536	540
Total		70	1,230	1,300

Prevalence of 5.4% (70/1,300) reflects prevalence in the IU population.

Clinical Performance of the Test	
Sensitivity	94.3% (66/70)
Specificity	43.6% (536/1,230)

Risks (Absolute Risks)

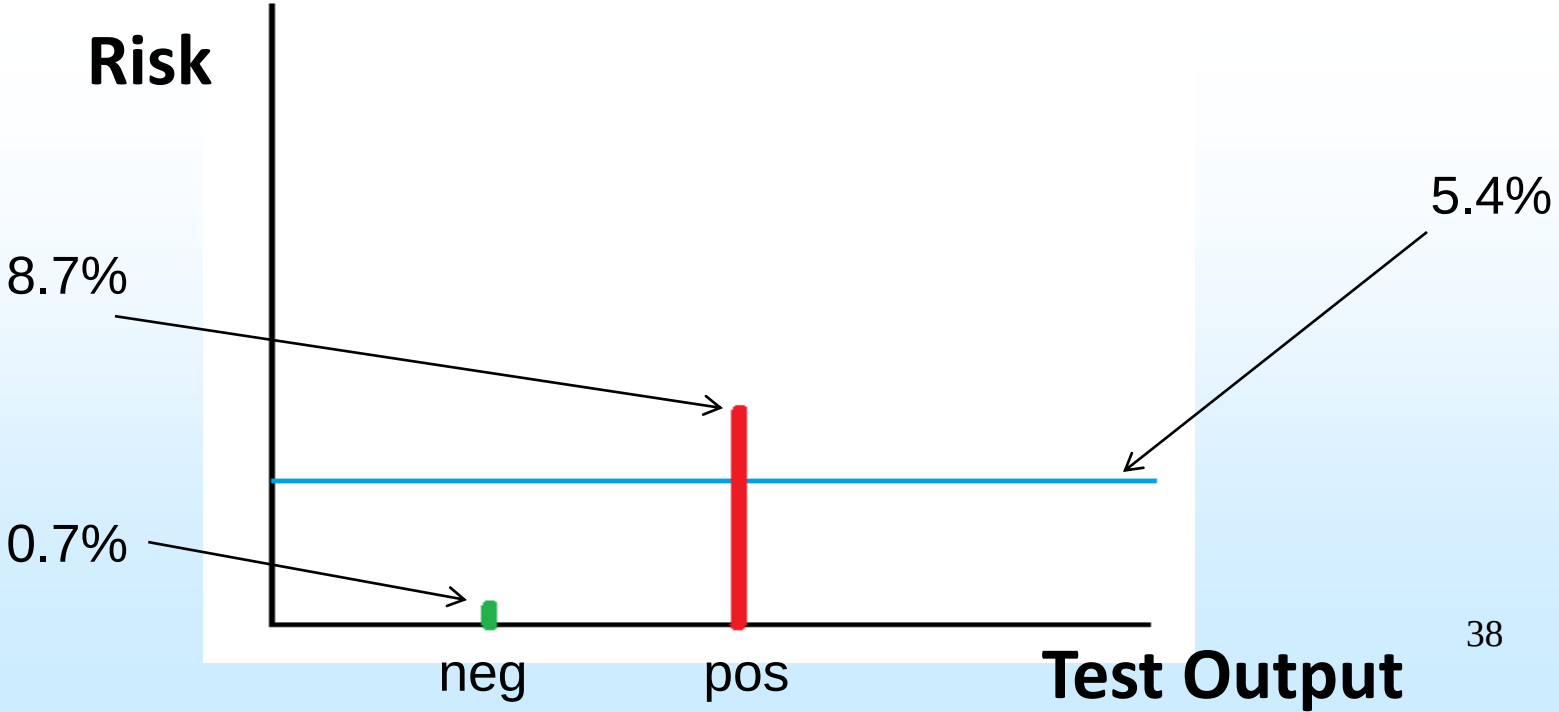
		D+	D-	Total
T	Pos	66	694	760
	Neg	4	536	540
Total		70	1,230	1,300

Clinical Performance of the Test	
R(Pos)=Risk of D+ for T pos (PPV)*	8.7% (66/760)
R(Neg)=Risk of D+ for T neg (1-NPV)*	0.7% (4/540)
π = Pre-test risk of D+ (baseline risk, prevalence)	5.4% (70/1,300)

*Post-test risk for T pos, post-test risk for T neg.

Absolute Risks

Clinical Performance of the Test	
R (Pos) =Risk of D+ for T pos	8.7% (66/760)
R (Neg)=Risk of D+ for T neg	0.7% (4/540)
π = Pre-test risk of D+	5.4% (70/1,300)



Consider Test with Multiple Outputs

Example #1: Multiplex test detecting two biomarkers A and B

These biomarkers are related to disease D

Four outcomes of the test:

(A+, B+)

(A+, B-)

(A-, B+)

(A-, B-)

Example #2: Test detects one biomarker (one SNP).

This biomarker is related to disease D.

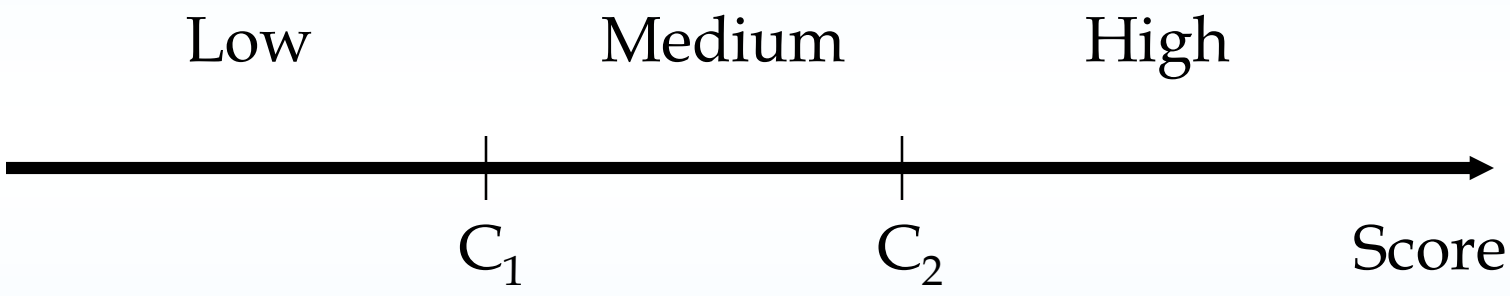
The biomarker has 3 possible results

(aa, aA, AA).

Example #3:

10 biomarkers combined in a score.

2 cutoffs are established that the score is reported as (High, Medium, Low)



How to describe performance of these tests?

Example : HPV Genotyping - 3 outcomes (HPV16/18); (Other High HPV types), (HPV neg)

Test Results	Colposcopy/Biopsy		Total
	CIN2+	Not-CIN2+	
HPV 16/18	46	314	360
Other HPV types	20	380	400
HPV neg	4	536	540
Total	70	1230	1300

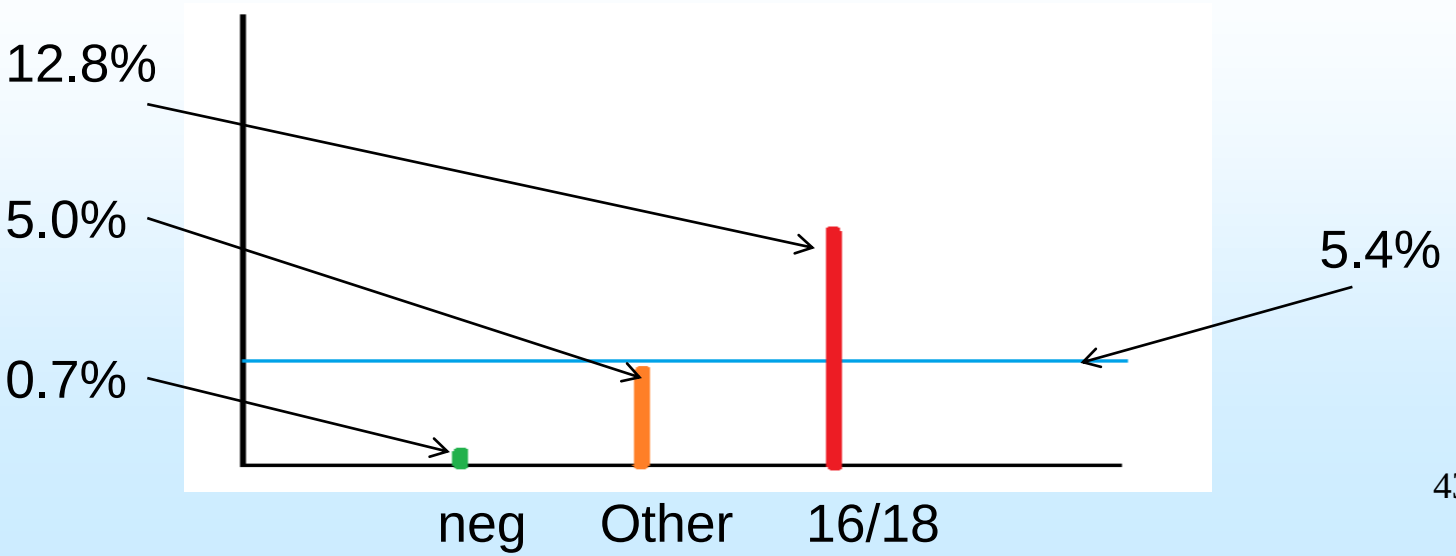
How to describe performance of this test?

Test with 3 outcomes: there are 3 risks R_X=Pr(D+|T=X)

Test Results	**Colposcopy/Biopsy**		Total	**Risk of CIN2+**
	CIN2+	**Not-CIN2+**		
HPV 16/18	46	314	360	**12.8%** **(46/360)**
Other HPV types	20	380	400	**5.0%** **(20/400)**
No HPV	4	536	540	**0.7%** **(4/540)**
Total	70	1230	1300	**5.4%** **(70/1300)**

Performance of the test is described by:
1) three risks; 2) three frequencies (percent) of results; 3) pre-test probability; 4) three likelihood ratios.

Test Results	Risk of Disease	Percent of results
HPV 16/18	12.8%	27.7%
Other HPV types	5.0%	30.8%
No HPV	0.7%	41.5%
Pre-test risk of CIN2+ is 5.4%		



Guidance for Industry and Food and Drug Administration Staff –

Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approval and De Novo Classifications (2016)

Example 3, pages 19-21

Thank you!

Marina.Kondratovich@fda.hhs.gov