Key Statistical Aspects of a Clinical Study

2014 Pre-Submissions Workshop

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

April 9, 2014

Outline

I. Introduction

II. Clinical performance characteristics:
    Risks, absolute risks, relative risks;
    Likelihood ratios (LR).

III. Potential Biases in Clinical Study

IV. Benefit-Risk Analysis

I. Introduction

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;
- **Monitoring:** is therapy working for a patient?;

* This is not a comprehensive list

Intended Use/Indication For Use

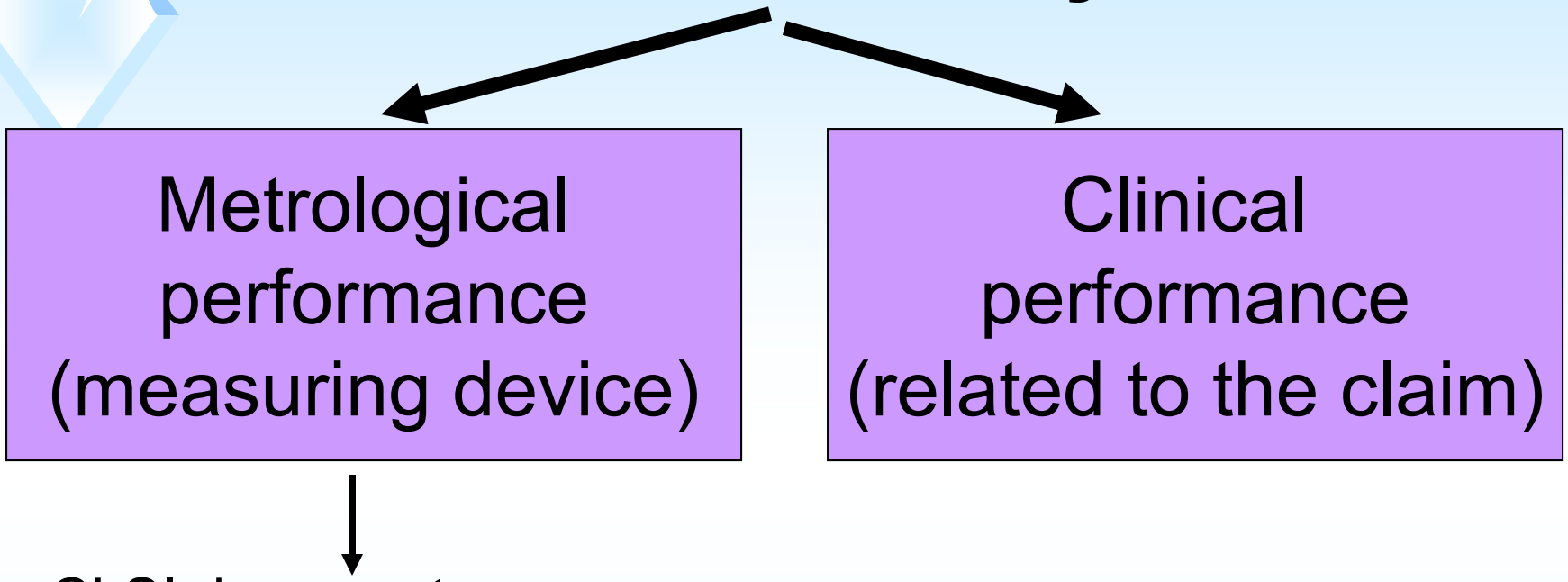
Example

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.*

- **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?

Medical Laboratory Test



CLSI documents helpful for analytical studies

- EP05-A2 Precision (under revision)
- EP06-A Linearity (under revision)
- EP07-A2 Interference Testing (under revision)
- EP09-A3 Systematic difference (bias)
- EP12-A2 Qualitative Test Performance
- EP17-A2 LoB, LoD and LoQ
- EP21-A Total error (accuracy) (under revision)
- EP25-A Stability of reagents
- EP28-A3c Reference intervals (old code C28-A3c)
- MM17-A Multiplex tests
- EP30-A Commutability (old code C53-A)
- EP32-R Traceability (under revision) (old code X05-R)

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

Clinical Studies

Typical scheme

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

Every subject

Candidate Test:	Assessment of Target Condition (ATC) (Gold Standard):
Positive, Negative	D+ = Target condition present, D- =Target condition absent

		Assessment of TC (Gold Standard)		Total
		Disease Present	Disease Absent	
Candidate Test	Pos	66	694	760
	Neg	4	536	540
Total		70	1,230	1,300

Clinical performance refers to the degree of agreement between the results of the Candidate test and the results from the Assessment of Target Condition (“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

Assessment of Target Condition

ATC (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.

Confusion may sometimes arise when distinguishing between:

- **Reference Method**

  related to analytical performance (best method for measuring of analyte (quantitative) or for detection of analyte (qualitative))

- **Assessment of Target Condition**

  related to clinical performance
  (no recognized term, other terms as “gold standard”, “reference standard”, “clinical reference standard”, “diagnostic accuracy criteria”)

- Most of the time, reference method and an assessment of the target condition are different (e.g., HPV test for cervical cancer, total PSA for the prostate cancer).
- Sometimes, reference method and ATC are identical (e.g., flu test).

Assessment of Target Condition

Basic principles:

1) Candidate test results CANNOT be used in the Assessment of Target Condition (ATC)

2) ATC can classify each subject from the target population as “Target condition present” (Disease present) or “Target condition absent” (Disease absent).

Archived (retrospective) samples

A good reason for pre-Sub

May be allowed

- 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 performance.

II. Clinical Performance Characteristics

Clinical Performance Characteristics

- Clinical sensitivity and clinical specificity
- Positive and negative predictive values along with prevalence
- Absolute risks and relative risks

Consider Test with Two Outcomes (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 a ATC ("Gold Standard") (D+, D-).

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

Clinical Performance of the Test	
R_1=Risk of D+ for T+ (PPV)*	8.7% (66/760)
R_0=Risk of D+ for T- (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 +, post-test risk for T -.

Absolute Risks, Relative Risks

Clinical Performance of the Test	
R_1=Risk of D+ for T+ (PPV)	8.7% (66/760)
R_0=Risk of D+ for T- (1-NPV)	0.7% (4/540)
π = Pre-test risk of D+	5.4% (70/1,300)

- $\mathbf{R_1/\pi = 1.6}$: For a subject with T+, the risk increases by 1.6 times with regard to pre-test risk (=8.7%/5.4%);
- $\mathbf{R_0/\pi = 0.14}$: For a subject with T-, the risk increases by 0.14 times (decreased by 7.3 (1/0.14) times) with regard to pre-test risk (=0.7%/5.4%);
- $\mathbf{R_1/R_0 = 11.7}$: For a subject with T+, the risk increases by 11.7 times with regard to the subjects with T- (=8.7%/0.7%)

Test with Multiple Outcomes

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 four biomarkers (four SNPs).

These biomarkers are related to disease D.

Each biomarker has 3 possible results

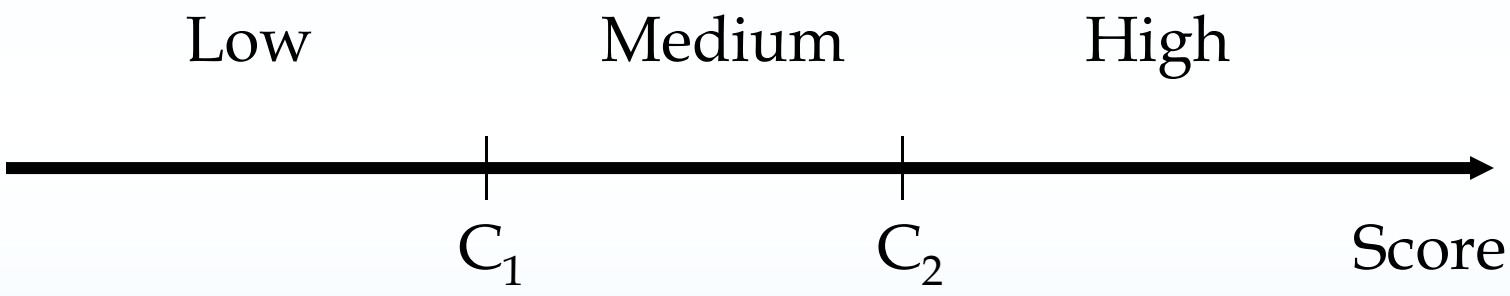
(aa, aA, AA).

Then test has 81 possible results: 81=3x3x3x3.

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)

We can calculate 3 likelihood ratios LR(T=X)

R_X depends on LR(T=X) and prevalence

$$\frac{R_X}{1-R_X} = LR(T=X) \times \frac{\pi}{1-\pi}$$

Test Results	Colposcopy/Biopsy		LR
	CIN2+	Not-CIN2+	
HPV 16/18	46 65.7%	314 25.5%	**2.6** (65.7%/25.5%)
Other HPV types	20 28.6%	380 30.9%	**0.93** (28.6%/30.9%)
No HPV	4 5.7%	536 43.6%	**0.13** (5.7%/43.6%)
Total	70 100%	1230 100%	

Performance of the test with three outcomes is described: 1) pre-test probability; 2) three LRs; 3) three frequencies (percent) of results.

Test Results	LR	Risk of Disease	Percent of results
HPV 16/18	2.6	12.8%	27.7%
Other HPV types	0.93	5.0%	30.8%
No HPV	0.13	0.7%	41.5%
Pre-test probability of CIN2+ is 5.4%			

Summary

- Clinical performance of the qualitative test with two outcomes can be described by a pair of sensitivity and specificity, or by PPV, NPV and prevalence.

- Risks and relative risks measure probabilities of events in a way that is interpretable and consistent with how people think.

- In addition, clinical performance of a medical test can be described by likelihood ratios.

Summary

Advantages:

- LRs do not depend on the prevalence;
- Absolute risks depend on the corresponding LR and prevalence;

- LRs are useful for comparing two qualitative tests with binary outcomes;
- LRs are useful for describing performance of the tests with multiple outcomes.

Because they do not depend on the pre-test risk, LRs and ORs can be calculated even in the case-control studies.

It is easy to adjust an OR for other variables (logistic regression)

III. 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** (when the study population does not represent the IU population)
2) **Spectrum bias**
3) **Verification bias**

1) Selection Bias

Examples of inappropriate study design

- 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

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*

❑ ***We know that some assessments of target condition are expensive or invasive: it may be impossible, or even unethical, to apply the ATC to <u>all</u> clinical study subjects.***

Examples

• Claims related to screening;

• Test under investigation is applied to in vitro samples and the ATC is applied to human subjects.

3) *Verification Bias*

Example

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

		ATC		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 ATC based on the “Current clinical practice”.

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

Analysis of the data with verified disease status

		ATC		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)

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 (2012)

The web address

http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm267829.htm

Example 3, pages 17-19

Hypothetical Example

Intended use population: women with mammogram results of BI-RADS 4;
Current practice – immediate biopsy is recommended
Prevalence of breast cancer is 25%

NEW Test: Negative – immediate biopsy is not recommended, wait a few months for further tests
Positive – immediate biopsy is recommended

		Biopsy		
		Malignancy	Benign	
Test	Positive	97	75	172
	Negative	3	225	228
		100	300	400

Performance:

Sensitivity= 97% (97/100) with 95% two-sided CI: 91.5% to 99.0%
Specificity = 75% (225/300) with 95% two-sided CI: 69.8% to 79.6%
Prevalence=25% (100/400)
Percent of positive result= 43.0% (172/400)
Risk of Malignancy for Test Pos = 56.4% (97/172)
Risk of malignancy for Test Neg = 1.3% (3/228)

Hypothetical Example

Consider 1,000 patients
Using performance data – recreate a table of performance

		Biopsy		
		Malignancy	Benign	
Test	Positive	243	187	430
	Negative	7	563	570
		250	750	1000

	Benefits		**Risks**	
	TP	TN	FN	FP
Cur. Pract.	250	0	0	750
New Test	243	563	7	187

Additional factors: uncertainty, patient tolerance for risk and perspective of benefit, availability of alternative diagnostics; risk mitigation.

Always consider benefits and risks SEPARATELY.

Thank you!

Marina.Kondratovich@fda.hhs.gov