FDA's "Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests": An Interactive Session

AMDM/FDA– OIVD 510(k) WORKSHOP
April 20-21, 2010

Kristen Meier, Ph.D.
Mathematical Statistician, Division of Biostatistics
Office of Surveillance and Biometrics
Center for Devices and Radiological Health, FDA

Final Guidance

- Final issued on March 13, 2007

http://www.fda.gov/cdrh/osb/guidance/1620.pdf

- DRAFT Guidance issued on March 12, 2003

Intent of Guidance

- help manufacturers and FDA reviewers
- describes information FDA needs in diagnostic device submissions for more efficient FDA review
- encourage use of standard terminology to provide clear, accurate and informative labeling for users
- identify common reporting mistakes that should be avoided

Statistical Guidance Scope

- for *all* diagnostic products not just *in vitro* diagnostics
- focus on diagnostic devices with 2 possible outcomes (positive/negative)
- *general concepts apply to <u>any</u> kind of diagnostic device*
  - importance of matching study design with intended use
  - clear data accounting and reporting results
  - minimize bias (internal validity)
  - desire for generalizability (external validity)

Use of Medical Tests

- Medical tests regulated by use, not just technology
- Use helps determine
  - path to approval
  - clinical questions and statistical hypotheses
  - study design
- One big message in guidance - describe test use

Same Technology - Different Uses

Example: uses of human papillomavirus (HPV) DNA test

- Measure specific HPV DNA types
  - Quantitative signal accurate?
- Screen women with abnormal PAP to determine need for colposcopy
  - Test+ correctly identify women w/abn PAP who need colposcopy?
- Assess presence/absence of high risk HPV types associated with cervical cancer in women over 30
  - Test– correctly identify women without cervical cancer?
- Outcome study
  - does clinical use of the test reduce cervical cancer deaths?

Diagnostic Intended Use (IU)
(how/by whom device is used)

- What is the device measuring, identifying or detecting?
  - analyte, organism, clinical condition
- What type of data output?
  - quantitative, semi-quantitative, qualitative
- Specimen type(s), source(s), matrix(-ces)
- Conditions for use?
  - hospital lab, physician's office, home use, …

Diagnostic Indications for Use (IFU)
(for what/on whom device is used)

- *target condition* (condition of interest)
  - a particular disease, a disease stage, health status, or any other identifiable condition or event within a patient, or a health condition that should prompt clinical action such as the initiation, modification or termination of treatment
- *intended use population* (target population)
  - those subjects/patients for whom the test is intended to be used
  - examples: general population (screen), subjects with particular signs and symptoms, pediatrics

Example IU

The Recent Respiratory Viral Panel (RRVP) is a qualitative new technology multiplex test intended for the simultaneous detection and identification of multiple respiratory virus nucleic acids in nasopharyngeal swabs from individuals suspected of respiratory tract infections. The following virus types and subtypes are identified using RRVP: Influenza A, Influenza A subtype H1, Influenza A subtype H3, ……

The detection and identification of these analytes from individuals exhibiting signs and symptoms of respiratory infection aids in the diagnosis of respiratory viral infection if used in conjunction with other clinical and laboratory findings. It is recommended that specimens found to be negative after examination using RRVP be confirmed by cell culture. Negative results do not preclude respiratory virus infection and should not be used as the sole basis for diagnosis, treatment or other management decisions.

Positive results do not rule out bacterial infection, or co-infection with other viruses. The agent detected may not be the definite cause of disease. The use of additional laboratory testing (e.g. bacterial culture, immunofluorescence, radiography) and clinical presentation must be taken into consideration in order to obtain the final diagnosis of respiratory viral infection.

Scope of Guidance

Diagnostic tests *for detection*

- analyte present (+) or absent (−)
- analyte level ≥ cutoff, or
  analyte level < cutoff
- clinical condition present (+) or absent (−)

Guidance Considers "Simplest" Case

	Truth	
	+	–
New +	44	1
Test –	7	168
Total	51	169

This is not so simple!

Statistical Guidance Developed

- what constitutes “truth”?
- what to do if we don’t know “truth”?
- what name do we give performance measures when we don’t have truth?
- what is the potential for *bias* and *heterogeneity* in device performance and *external validity* of study results? (do the study and subjects *represent* the IU and IFU population?)

Benchmarks for Assessing Diagnostic Performance

Move away from notion of "truth"

FDA recognizes 2 categories of benchmarks:

- ***(clinical) reference standard***
- ***non-reference standard*** (a method or predicate other than a reference standard; due to 510(k) regulations)

(Clinical) Reference Standard

- "considered to be the best available method for establishing the presence or absence of the target condition…it can be a single test or method, or a combination of methods and techniques, including clinical follow-up"
  (Bossuyt et al. 2003)
- does not consider outcome of new test under evaluation (see *discrepant resolution* in guidance)

Reference Standard (FDA)

What constitutes "best available method"/reference standard?

- opinion and practice within the medical, laboratory and regulatory community
- several possible methods could be considered
- maybe no consensus reference standard exists
- maybe reference standard exists but for non-negligible % or intended use population, the reference standard is known to be in error
- *will evolve over time!*

Not a statistical call, but statistical principles can help

Choice of Reference Standard

- driven by IFU (target condition and intended use population)
- if multiple IU and IFUs then each needs supporting evidence/data

Example of Reference Standard

Candidate device: human papillomavirus (HPV) DNA test for cervical cancer

Clinical Reference Standard: diagnosis of cervical cancer determined by a specified algorithm combining results of cytology, histology, HPV DNA from non-candidate method, and clinical follow-up.

Analytical concerns: HPV DNA test is well calibrated (as determined by *reference method*) and precise

Example of Reference Standard

- *Test:* human papillomavirus (HPV)
- *Reference standard:* Patient disease status based on cytology, colposcopy and histopathology of the cervical biopsy according to the table below. A condition of interest is HSIL or greater disease.

Cytology Result	*Histology Result*	*Disease Status*
NEG	NEG or ND*	NEG
LSIL	NEG	LSIL
HSIL	NEG	HSIL
Cancer	NEG	HSIL+
NEG	LSIL	LSIL
LSIL	ND*	LSIL
LSIL	LSIL	LSIL
HSIL	LSIL	LSIL
Cancer	LSIL	LSIL
NEG	HSIL	HSIL
LSIL	HSIL	HSIL
HSIL	HSIL	HSIL
HSIL	ND*	HSIL
Cancer	HSIL	HSIL
NEG	Cancer	HSIL+
LSIL	Cancer	HSIL+
HSIL	Cancer	HSIL+
Cancer	ND*	HSIL+

*Biopsy and/or ECC not done because no abnormalities were observed upon colposcopy or histology result not available

Choosing a Reference Standard

- Consult with FDA about what is an appropriate reference standard *before* starting your study

- What do you do if there is no reference standard or it is impractical to use on all subjects (e.g., autopsy., biopsy)?

Choosing a Comparative Benchmark

- If reference standard is available – use it
- If reference standard is available but impractical – use it to the extent possible (requires complex statistical design and analysis)
- If reference standard is not available
  - construct one
  - use a non-reference standard

Choice of Benchmark

► ***Use terminology appropriate for your benchmark***

Clinical Reference Standard

- report sensitivity, specificity, predictive values of positive and negative results, likelihood ratios
- terms from scientific literature

Non-reference standard

- report *positive percent agreement* and *negative percent agreement* (do not use *relative* sens/spec)
- FDA created terms to address 510(k) regulations

Test Performance: Dichotomous Test

		Study Population TRUTH	
		Truth+	Truth–
New Test	Test+	TP (true+)	FP (false+)
	Test–	FN (false–)	TN (true–)

sensitivity (sens): 100%×TP/(TP+FN)

specificity (spec): 100%×TN/(FP+TN)

Useful for interpretation (depends on prevalence):

positive predictive value (PPV): 100%×TP/(TP+FP)

negative predictive value (NPV): 100%×TN/(FN+TN)

Example: Estimating Sensitivity and Specificity

		Reference Standard +	Reference Standard –
New	+	44	1
Test	–	7	168
Total		51	169

Sensitivity (sens): 100%×44/51 = 86.3%

Specificity (spec): 100%×168/169 = 99.4%

"Perfect" Test

sensitivity = specificity = 100%

		Reference Standard	
		+	–
New	+	51	0
Test	–	0	169
Total		51	169

"Useless" Test

sensitivity = 100% – specificity

		Reference Standard +	Reference Standard –
New	+	46	152
Test	–	5	17
Total		51	169

sens = 90% (46/51)

100% – spec = 90% (152/169)

Agreement (to non-reference standard)

		Non-Reference Standard	
		+	–
New Test	Test+	a	b
	Test–	c	d

PPA: Positive percent agreement (new/non ref. std.)
$=100\% \times a/(a+c)$

NPA: Negative percent agreement (new/non ref. std.)
$=100\% \times d/(b+d)$

Commonly reported, but not very useful by itself:

Overall agreement $= 100\% \times (a+d)/(a+b+c+d)$

Agreement - Example

		Study Population Non-Reference Standard	
		+	–
New	+	40	5
Test	–	4	171
Total		44	176

Positive percent agreement (PPA) = 90.9% (40/44)
Negative percent agreement (NPA) = 97.2% (171/176)

Same arithmetic as calculating sens and spec, but interpretation is very different!

Interpretation
Sens/spec vs. Agreement

- If sens=spec =100%, then the new test is "perfect"

- Is it desirable to have PPA=NPA=100%?

Agreement

- has value in supporting substantial equivalence (SE)
- agreement is *not* accuracy

  agreement ≠ "correct"
- see Guidance Appendix for pitfalls of agreement measures
- best to have 3-way comparison data between the new test, the predicate and a reference standard

Bias (in Performance Estimates)

- a concern regardless of benchmark used
- biased performance estimates are systematically too high or too low
- can arise due to type study design or data analysis
- often can't quantify bias
- to help reduce bias get the *right* data, not necessarily *more* data

Sources/Types of Bias: AVOID!

- comparative benchmark has error
- reference standard uses outcome of candidate test
- study does not include the "right" subjects (*spectrum effect)*
  - subjects not in IU population
  - only extreme cases included
- non-representative subset of subjects evaluated by reference standard, no statistical adjustments made to estimates (*verification* or *work-up bias)*
- revise comparative data and performance estimates based on discrepant resolution
- discard equivocal results (*reporting bias)*

Discrepant Resolution - Avoid

- problematic attempt to adjust performance measures for error in the benchmark
  - when the new device and the benchmark results agree, assume both are correct
  - when they disagree, retest the subject using a third test and change the benchmark result to the retest result
  - "agreement" always increases or stays the same
- procedure does not adjust for benchmark error and may add additional bias to performance estimates
- see Guidance Appendix for more details

Do Not Exclude "Equivocals"

When test has an intermediate or equivocal zone in between positive and negative…

- report all results as a 2×3, 3×2 or 3×3 table
- to calculate PPA and NPA make a 2×2; combine (dichotomize) results into two categories:
  - {positive and equivocal} versus negative
  - positive versus {equivocal and negative}

Make a 2×2: Dichotomize Results

Comparative Method (CM)

New Test	+	Eq	−
+	40	1	3
Eq	0	2	1
−	4	3	121

43	4
7	121

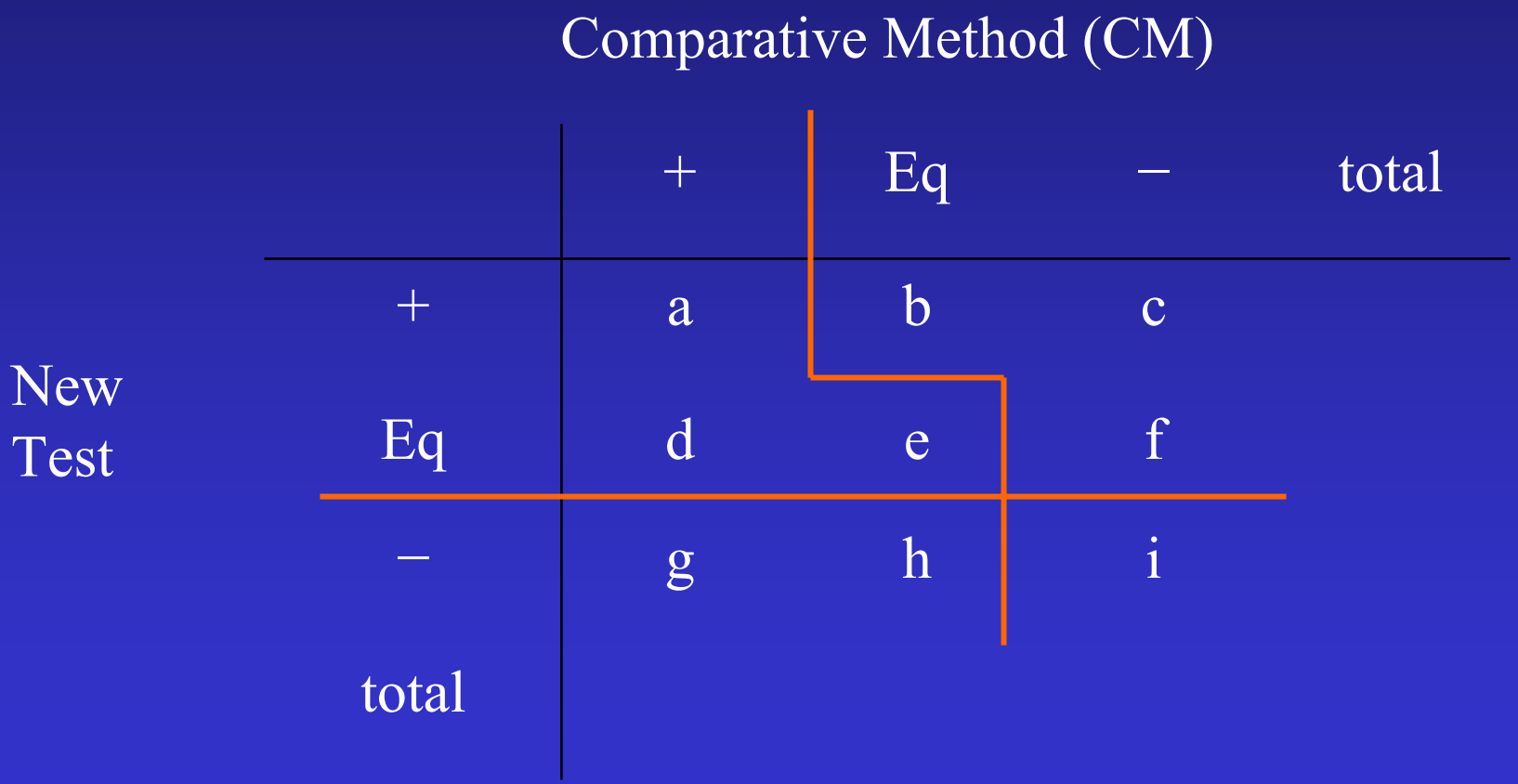
Make a 2×2: Dichotomize Results

		Comparative Method (CM)		
	+	Eq	–	total
New Test +	a	b	c	
New Test Eq	d	e	f	
New Test –	g	h	i	
total				

Make a 2×2: Dichotomize Results

	Comparative Method (CM)			
New Test	+	Eq	−	total
+	a	b	c	
Eq	d	e	f	
−	g	h	i	
total				

Inappropriate Combining for 2×2



Do Not Discard "Equivocals"

- more than one way to combine results
  - what makes sense clinically; how are patients managed?
  - OK to report more than one set of PPA and NPA
- do not use outcome of new test to decide how to dichotomize the comparative method
- Alternative? report percent agreement for each category of the comparative method

General Practices to Avoid

Do Not:

- use terms “sensitivity” and “specificity” if reference standard is not used
- use test under evaluation in diagnostic workup or to establish diagnosis
- use data altered or updated by discrepant resolution
- discard equivocal results in data presentations and calculations

Good Practices (External Validity)

Do:

- include appropriate subjects and/or specimens (per IU and IFU)
- use final version of the device according to the final instructions for use
- use several of these devices in your study
- include multiple users/operators with relevant training and range of expertise
- cover a range of expected use and operating conditions
- see “Reporting Recommendations” in guidance (Section 5, pages 14-17)

"Reporting Recommendation" Highlights

- report 2×2 table of results
- sens, spec, reference standard and condition of interest is a package deal – report it all!
- describe the study population (on whom and by whom device is used in study)
- if reference standard not used, report results as PPA and NPA
- report equivocal (gray zone) results and invalid results (device fails built in controls or fails to give a result)
- report all percentages as fractions
  - example: estimated sens is 96.9% (94/97)

STARD Initiative

STAndards for ***R***eporting of ***D***iagnostic Accuracy Initiative (pronounced STAR-D)

- effort by international working group (academia, government, clinical laboratories)
- goal: "to improve the accuracy and completeness of reporting of studies of diagnostic accuracy, to allow readers to assess the potential for bias in the study (internal validity) and to evaluate its generalizability (external validity)"
- checklist of 25 items to include when reporting results
- provide definitions for terminology
- recommendations adopted in over 200 biomedical journals
- http://www.stard-statement.org

Download it and read it!

Conclusions

- correct terminology & complete reporting is important for safe & effective use of device
- this guidance can be a very useful tool and includes good references in bibliography
- many concepts apply to *any* diagnostic device
- consult with FDA when planning your study

References

See bibliography in Guidance

Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry*. 2003;49(1):1-18. (see also http://www.stard-statement.org)

CLSI. *User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline—Second Edition*. CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.


Center for Devices and Radiological Health
CDRH