

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

AMDM/FDA– OIVD 510(k) WORKSHOP
April 21-22, 2009

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- DRAFT Guidance issued on March 12, 2003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 (as in STARD) to provide clear, accurate and informative labeling for users
- identify common reporting mistakes that should be avoided

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!

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 any 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)

Regulatory Perspective: “Diagnostic Device” – a Package Deal

- package is the combination of the physical device or software, the *intended use* (how/by whom the device is used) and *indications for use* (for what/on whom device is used)
- changing one thing in the package can change the device in terms of performance, safety and effectiveness so that a new submission required

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

Scope of Guidance

- *target condition* is present (+) or absent (–)
- yes/no or +/ – device
 - inherent
 - dichotomize with a cut-off

Dichotomize with Cut-off

quantitative output



result < cut-off

result \geq cut-off

—

+

Negative

Positive

Guidance Considers “Simplest” Case

	Truth	
	+	-
New +	44	1
Test -	<u>7</u>	<u>168</u>
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:

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

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”
- 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 calibrated and precise

Example with no Reference Standard

(This example not within scope of Guidance document)

Candidate device: test to detect human papillomavirus (HPV) DNA

- quantitative IFU, not yes/no
- need a Reference *Method* to evaluate analytical performance (including trueness and precision) of HPV DNA test

Terminology Note - Reference Standard

- ISO definition of “reference standard” (standard having highest metrological quality...) is different
- reference *standard* is usually different from reference *method* or reference *material* (for use in calibration)
- CLSI EP12-A2 uses term “diagnostic accuracy criteria” for reference standard concept

see Clinical Laboratory Standards Institute’s
Harmonized Terminology Database at www.clsi.org

Example Reference Standard

Test: Enzyme immunological assay for the measurement of Type 1 collagen C-Telopeptides in plasma and serum.

- IFU: Used as an aid in predicting skeletal response (Bone Mineral Density or BMD) in post menopausal women under going anti-resorptive therapies (HRT and biphosphonate therapies).

Ref Standard/condition of interest: change in bone mineral density (BMD) \leq 1 versus BMD $>$ 1 where BMD is measured at the lumbar spine using dual-energy x-ray densitometry and change is defined as the slope of the linear regression line of spine BMD versus time over a 3-year period.

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*

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 23

Test Performance: Dichotomous Test

Study Population

TRUTH

		<u>Truth+</u>	<u>Truth-</u>
New	Test+	TP (true+)	FP (false+)
Test	Test-	FN (false-)	TN (true-)

sensitivity (sens): $100\% \times TP / (TP + FN)$

specificity (spec): $100\% \times TN / (FP + TN)$

Useful for interpretation (depends on prevalence):

positive predictive value (PPV): $100\% \times TP / (TP + FP)$

negative predictive value (NPV): $100\% \times TN / (FN + TN)$

Example: Estimating Sensitivity and Specificity

Reference Standard

		+	-
New Test	+	44	1
	-	<u>7</u>	168
Total		51	169

Sensitivity (sens): $100\% \times 44/51 = 86.3\%$

Specificity (spec): $100\% \times 168/169 = 99.4\%$

Estimating Sensitivity and Specificity

- “Perfect” test: $\text{sens}=\text{spec}=100\%$

	Reference Standard	
	+	-
New	+	51 0
Test-		<u>0 169</u>
Total		51 169

- “Useless” test: $\text{sens}=100\%-\text{spec}$

	Reference Standard		
	+	-	
New	+	46 152	$\text{sens}=90\% (46/51)$
Test-		<u>5 17</u>	$\text{spec}=10\% (17/169)$
Total		51 169	$1-\text{spec}=90\% (152/169)$

Agreement

Non-Reference Standard

		+	-
New Test	Test+	a	b
Test	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	
		+	-
New Test	+	40	5
	-	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 $\text{sens}=\text{spec} = 100\%$, then the new test is “perfect”
- Is it desirable to have $\text{PPA}=\text{NPA}=100\%$?

Agreement

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

- biased performance estimates are systematically too high or too low
- can arise due type study design or data analysis
- a concern regardless of benchmark used
- 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*)

Example of Inappropriate Ref. Standard

- *Test:* Chlamydia trachomatis (CT) and Neisseria Gonorrhoeae (GC) amplified DNA test for urine, endocervical swabs, and male urethral swabs
- *Ref Standard:* results from culture, marketed DNA amplification assay (predicate), Direct Fluorescent Antibody (DFA) test
 - inappropriate ref standard positive: cell culture +; or, DFA + on specimens that are culture–/new test + or culture–/predicate–
 - appropriate ref standard positive: cell culture +; or, DFA + on specimens that are culture –/predicate–

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”

Guidance: report 2 sets of performance counting equivocals as +, then as –

Not in Guidance: better way for *how to include* equivocal results depends on...

- characteristics of the test
 - underlying quantitative signal that is linear, well-calibrated?
- rationale for an equivocal zone based on
 - imprecision of quantitative signal near cutoff
 - analyte not the best discriminator between condition present/absent
 - analyte status is changing in short time period (“seroconversion”)

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)

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.

