Considerations for Evaluation of Diagnostic Performance

Felipe A. Medeiros, M.D., Ph.D.
Professor of Ophthalmology
Hamilton Glaucoma Center
University of California, San Diego
I have the following financial disclosures:

• Research Support:
  – National Eye Institute R01 EY021818
  – Alcon, Inc.
  – Allergan, Inc.
  – Carl-Zeiss Meditec, Inc.
  – Heidelberg Engineering
  – Optovue
  – Reichert
  – Sensimed
  – Merck, Inc.
Potential Applications of Imaging Instruments

• **Diagnosis**
  - Decrease diagnostic uncertainty in those **SUSPECTED** of a condition (suspected of having damage or suspected of having progression)

• **Screening**
  - Identify abnormal or suspected cases in the general population or pre-selected subjects (e.g., older population, positive family history)

• **Prognosis**
  - Determine risk of developing a condition
The design of the study should take into account the purpose of the test and involve the clinically relevant population.
Diagnostic Accuracy Studies

• Diagnostic tests are used to decrease uncertainty about presence of a condition

• **Cross-sectional assessment**: The test is applied at a single point in time in those suspected of having the disease

  *Does this patient have glaucoma?*

• **Longitudinal assessment**: The test is applied multiple times during follow-up in suspects or those with confirmed disease

  *Does this patient have disease progression?*
Uncertainty is what characterizes glaucoma suspects

Suspicious optic disc appearance, with normal or suspicious visual field results

Does this patient have glaucoma?
Diagnostic Studies in Imaging

• Diagnostic Studies in Glaucoma
  – **Cases:** Glaucoma patients with **repeatable visual field loss**
  – **Controls:** Healthy individuals (volunteers without any suspicious finding)

• Diagnostic accuracy measures
  – Proportion of cases correctly identified by the test as being abnormal (sensitivity)
  – Proportion of controls correctly identified by the test as being normal (specificity)
  – ROC curves
Typical cases included in most studies...

Why do we need an imaging instrument to diagnose glaucoma in this patient?
A typical control... healthy volunteer

IOP = 10 mmHg
Limitations of “conventional” studies

• Case-control studies including only patients with well-defined disease and healthy subjects
  ... are important for an initial evaluation of a diagnostic test

• However...

  ... In clinical practice, diagnostic tests are used to evaluate patients who are suspected of having the disease, not patients with confirmed disease

“when the diagnosis is obvious to the eye, we don’t need further diagnostic tests”
(Straus SE et al. Evidence-Based Medicine: How to Practice and Teach it)
Limitations of “conventional” studies

- The population sample included in most studies may not be representative of the one in which we apply the diagnostic tests in everyday practice.

- Estimates of sensitivity and specificity obtained from these studies may not be directly applicable in clinical practice.
Do test results distinguish patients with and without the target disorder among those in whom it is clinically sensible to suspect the disorder?

“If sensitivity is determined in seriously ill subjects and specificity in clearly healthy individuals, both will be grossly overestimated”

(Sackett)
Healthy sample, recruited from the general population

Test more abnormal (larger cup area or smaller rim area)
Healthy sample, recruited from the general population

Test more abnormal (larger cup area or smaller rim area)

Cut-off to determine abnormality
Only 1% false-positives
For example, abnormal MRA

Should I expect the test to have only 1% false positives in my clinical practice?
Healthy sample, recruited from the general population

Test more abnormal (larger cup area or smaller rim area)

Clinical population will be enriched by individuals with suspicious discs. The test will not have the same specificity
If the purpose of the test is to complement current clinical evaluation... How should we design diagnostic accuracy studies?
An example:

Accuracy of ECG for diagnosing myocardial infarction
Accuracy of ECG for myocardial infarction

- Review of diagnostic accuracy studies of ECG¹
- All included patients were suspect of MI at the time of the ECG
- Diagnosis of MI was subsequently confirmed or ruled out based on levels of cardiac enzymes (reference test)

---

Diagnostic Accuracy of ECG

Patients suspected of MI (Chest pain)

- Myocardial infarction
- Without myocardial infarction (Chest pain from other causes)

Reference test: Cardiac enzymes

How should we design diagnostic accuracy studies in glaucoma?
Diagnostic Accuracy of Test X

Patients suspected of having glaucoma (High IOP, suspicious discs, etc)

Glaucoma

No glaucoma

Reference test

TEST X
What reference standard should be used?

Visual Field – Not a good reference standard if you want to evaluate additional clinical benefit of test X
What reference standard should be used?

*Cross-sectional* optic disc evaluation (photos, slit-lamp)

Not a good reference standard because of poor accuracy, prognostic value
What reference standard should be used?

Diagnostic studies involving glaucoma suspects will require longitudinal follow-up (historical or prospective).
Is this a good design?

This is a **prognostic** study

Important to assess measures of prognostic ability (hazard ratios not enough)

Need to account for other risk factors

Test may have a weak prognostic ability but be a good diagnostic test

---

**SUSPECTS**

Follow them over time

Visual fields, photos

---

Progress:

Glaucoma

Do not progress

---

Apply the diagnostic test

**BASELINE**
Test may have a weak prognostic ability but be a good diagnostic test.

Ocular hypertensive eyes WITHOUT nerve damage

Progressive nerve damage: Glaucoma

Follow them over time

Imaging tests will not be able to discriminate these eyes at baseline (they don’t have damage) Poor prognostic value

Imaging tests may clearly separate eyes with nerve damage from those without damage here Good diagnostic value
Test may have a weak prognostic ability but be a good diagnostic test.

Ocular hypertensive eyes WITHOUT nerve damage

Progressive nerve damage:
- Glaucoma

No progression

Follow them over time

How can we assess diagnostic performance here if we cannot use visual fields as reference standard?

Historical follow-up
Progressive Optic Disc Damage is Highly Predictive of Development of Functional Loss in Glaucoma

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$R^2$ (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optic Disc Progression</td>
<td>79% (65% - 87%)</td>
</tr>
<tr>
<td>Baseline GON grading</td>
<td>21% (9% - 37%)</td>
</tr>
<tr>
<td>Baseline vertical C/D ratio</td>
<td>21% (8% - 37%)</td>
</tr>
<tr>
<td>Baseline IOP</td>
<td>10% (2% – 22%)</td>
</tr>
<tr>
<td>CCT</td>
<td>6% (1% – 15%)</td>
</tr>
<tr>
<td>Baseline PSD</td>
<td>26% (15% - 40%)</td>
</tr>
<tr>
<td>Age</td>
<td>23% (11% - 39%)</td>
</tr>
</tbody>
</table>

Does it matter?
The Effects of Study Design and Spectrum Bias on the Evaluation of Diagnostic Accuracy of Confocal Scanning Laser Ophthalmoscopy in Glaucoma

Felipe A. Medeiros, Diana Ng, Linda M. Zangwill, Pamela A. Sample, Christopher Bowd, and Robert N. Weinreb

- Influence of the studied population on the Diagnostic Accuracy of CSLO (HRT)

The Effect of Study Design...

- Compared Diagnostic Accuracy of CSLO in 2 scenarios:

  - **ANALYSIS 1**
    - Discriminate patients with glaucomatous visual field loss from healthy subjects (recruited from general population)

  - **ANALYSIS 2**
    - Discriminate suspects who have glaucoma from suspects who do not have glaucoma (using history of previous optic disc progression as reference standard)

Analysis 2
Cohort of Glaucoma Suspects

Suspect A

Suspect B
Analysis 2

Patient followed for 18 years without treatment and without any changes to the optic nerve and VF
Evidence of PROGRESSIVE glaucomatous damage confirms diagnosis of glaucoma
The Effect of Study Design...

- Differences in diagnostic accuracy in the two analyses
  - Glaucoma Probability Score (GPS)

The Effect of Study Design...

Analysis 1
ROC curve area = 0.89

Analysis 2
ROC curve area = 0.65

Spectrum Bias

- If patients are referred as suspicious of glaucoma by the same characteristic measured by the diagnostic test (e.g., rim thinning), the test will have little additional value for decreasing diagnostic uncertainty
Potential Applications of Imaging Instruments

- **Diagnosis**
  - Decrease diagnostic uncertainty in those SUSPECTED of a condition (suspected of having damage or suspected of having progression)

- **Screening**
  - Identify abnormal or suspected cases in the general population or pre-selected subjects (e.g., older population, positive family history)

- **Prognosis**
  - Determine risk of developing a condition
Screening

• Estimates of diagnostic accuracy obtained in clearly glaucomatous versus healthy eyes seem to be more relevant to opportunistic screening situations
Diagnostic Studies
SDOCT Retinal Nerve Fiber Layer

Felipe A. Medeiros, M.D., Ph.D.
Professor of Ophthalmology
Hamilton Glaucoma Center
University of California, San Diego
• 134 glaucoma suspects followed for average of 14 ± 3.6 years
  • 42 progressive GON (photos)
  • 86 control eyes (no progression followed untreated)
RNFL parameters performed significantly better than topographic disc parameters for diagnosing damage in suspects.

ROC area SDOCT average thickness = 0.86

ROC area HRT rim area = 0.72
RNFL versus optic disc topography

- In clinical practice, patients are usually referred as suspected of having glaucoma because of suspicious rim thinning or large cups.

- In this situation, RNFL parameters seem to offer more benefit as a complementary test to clinical examination.
• 138 eyes with glaucomatous visual field defects
• 106 healthy eyes
• Diagnostic accuracies of the 3 SDOCT instruments
## Comparison of Cirrus HDOCT, RTVue and Spectralis

### Areas under ROC curves

<table>
<thead>
<tr>
<th></th>
<th>Spectralis</th>
<th>Cirrus</th>
<th>RTVue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global RNFL thickness</td>
<td>0.862</td>
<td>0.864</td>
<td>0.844</td>
</tr>
<tr>
<td>Inferior thickness</td>
<td>0.837</td>
<td>0.849</td>
<td>0.846</td>
</tr>
<tr>
<td>Superior thickness</td>
<td>0.865</td>
<td>0.863</td>
<td>0.841</td>
</tr>
</tbody>
</table>

Ophthalmology 2011;118:1334–1339
Diagnostic Studies of RNFL parameters

- Diagnostic accuracy estimates obtained from “conventional studies” (healthy versus glaucomatous field loss) are similar to those estimates obtained from cohort studies of suspects.
Disease severity still affects accuracy of RNFL parameters...
Effect of Disease Severity on the Performance of Cirrus Spectral-Domain OCT for Glaucoma Diagnosis

Mauro T. Leite,1,2 Linda M. Zangwill,1 Robert N. Weinreb,1 Harsha L. Rao,1,3 Luciana M. Alencar,1,4 Pamela A. Sample,1 and Felipe A. Medeiros1,2

VFI = 95%

VFI = 85%

IOVS 2010;51:4104–4109
### Areas under ROCs for different levels of disease severity (SDOCT)

<table>
<thead>
<tr>
<th>VFI</th>
<th>Global</th>
<th>Inferior</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>0.822</td>
<td>0.851</td>
<td>0.812</td>
</tr>
<tr>
<td>90%</td>
<td>0.886</td>
<td>0.871</td>
<td>0.874</td>
</tr>
<tr>
<td>80%</td>
<td>0.932</td>
<td>0.920</td>
<td>0.921</td>
</tr>
<tr>
<td>70%</td>
<td>0.962</td>
<td>0.954</td>
<td>0.952</td>
</tr>
</tbody>
</table>
Effect of Disease Severity on the Performance of Cirrus Spectral-Domain OCT for Glaucoma Diagnosis

Mauro T. Leite,1,2 Linda M. Zangwill,1 Robert N. Weinreb,1 Harsba L. Rao,1,3 Luciana M. Alencar,1,4 Pamela A. Sample,1 and Felipe A. Medeiros1,2
Does SDOCT RNFL assessment assist in glaucoma diagnosis?

“Diagnosis is not about finding the truth but limiting uncertainty”

(Straus SE et al. Evidence-Based Medicine: How to Practice and Teach it)
Evaluating a Glaucoma Suspect

Pre-test probability of disease = 30%
How can SDOCT help us with this patient?

Pre-test probability of glaucoma (30%)

Imaging Test

SDOCT result

Post-test probability of glaucoma?
How do we calculate the post-test probability?

Post-test odds of disease = LR \times Pre-test odds of disease

Odds = \frac{probability}{1 - probability}
Continuous Likelihood Ratios for Glaucoma Diagnosis with Spectral Domain OCT

The likelihood ratio for a specific test value is the tangent to the ROC curve at the corresponding value.

Lisboa R, Mansouri K, Zangwill L, Weinreb RN and Medeiros FA. ARVO 2011
Continuous Likelihood Ratios for Glaucoma Diagnosis with Spectral Domain OCT

Lisboa R, Mansouri K, Zangwill L, Weinreb RN and Medeiros FA. ARVO 2011
Evaluating a Glaucoma Suspect

07/15/2009

07/09/2009

Pre-test probability of disease = 30%  Post-test probability of disease = 93%

SDOCT result
Diagnostic Accuracy of SDOCT RNFL

- Diagnostic accuracy studies involving the clinically relevant population show a clear benefit of RNFL assessment with SDOCT in decreasing uncertainty in glaucoma diagnosis
Addressing the Questions Asked...
For a cross-sectional study describing the distribution of measurements in known normal and known diseased subjects (excluding suspects):

a) How should the diseased subject population be defined? What clinical work-up should be done to establish these populations?

A: Cross-sectional studies including only clearly disease subjects versus healthy normals may lead to biased estimates of accuracy. This seems to be worse for certain instruments/parameters versus others.

Studies involving healthy versus glaucomatous field loss should be conducted for initial evaluation, but should be followed by studies evaluating the clinically relevant population (suspects)
Can the diseased population be further divided by severity of disease using a clinical reference method (e.g., perimetry)? If so, should the clinical performance of the imaging device be characterized separately for each severity group?

A: YES. Important to include a broad spectrum of disease severity and use methods to evaluate diagnostic performance according to disease severity (e.g., ROC regression methods).
For a longitudinal diagnostic performance study where the structural measurement is taken at baseline and the clinical reference standard consists of a baseline assessment with follow-up:

a) What subjects should be included in the study population (including glaucoma suspects)?

This is a prognostic study. What are you trying to predict?

If we are trying to predict development of glaucoma, the population should be of suspects (without clear signs of damage)
For a longitudinal diagnostic performance study where the structural measurement is taken at baseline and the clinical reference standard consists of a baseline assessment with follow-up:

b) What is an appropriate clinical reference standard?

Again, depends on what you are trying to predict. If development of glaucoma, then appropriate standards could be development of visual field loss and/or progressive optic nerve damage.
For a longitudinal diagnostic performance study where the structural measurement is taken at baseline and the clinical reference standard consists of a baseline assessment with follow-up:

c) What minimum follow-up period would provide assurance that the subject has been appropriately classified?

Not necessarily we need to establish a minimum period, as long as we take into account censoring. It is very important to adequately evaluate prognostic ability (c-index, modified R2, reclassification tables, predictiveness measure)